The Dissertation Committee for Groves Bayne Dixon Certifies that this is the approved version of the following dissertation:

Genetic and Epigenetic Mechanisms of Adaptation in Stony Corals

Committee:

Mikhail V. Matz, Supervisor

Daniel I. Bolnick

Hans A. Hofmann

Vishy Iyer

Mark Kirkpatrick

Genetic and Epigenetic Mechanisms of Adaptation in Stony Corals

by

Groves Bayne Dixon

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August, 2017

Dedication

Dedicated to Jean Dixon.

Acknowledgements

Lots of people helped me with this. In particular, I would like to thank Sarah Davies, Rachel Wright, Marie Strader, my family, my committee, and Misha Matz.

Genetic and Epigenetic Mechanisms of Adaptation in Stony Corals

Groves Bayne Dixon, Ph.D.

The University of Texas at Austin, 2017

Supervisor: Mikhail V. Matz

In this dissertation, I used genomic techniques to examine interrelationships between genotype, gene expression, DNA methylation and environmental conditions in the model coral *Acropora millepora*. I present three major findings: 1) populations in the Great Barrier Reef have the potential for rapid genetic adaptation to climate change 2) patterns of DNA methylation predict gene expression plasticity 3) patterns of DNA methylation can predict fitness under environmental change.

Table of Contents

List of Tables

List of Figures

OVERVIEW

In a coral reef, trillions of genomes arrange themselves into complex, rock-making structures that alter the planet on a geologic scale. The most conspicuous example, 345,000 km^2 and visible from orbit, is the Great Barrier Reef off the east coast of Australia (De'ath et al. 2012). A goal of ecological genomics—the study of interactions between full sets of genetic material and their biotic and abiotic surroundings—is to understand how genomes create such complex systems; to describe how nucleotides produce reefs. A critical factor in such processes is adaptation, defined generally as the ability of an organism to survive and reproduce in its current environment. Adaptation comes about through genomic change: beneficial mutations which multiply as a result of natural selection, and chromatin modifications, such as DNA methylation, which can adaptively alter expression without changing the underlying genetic sequence. The major goal of this dissertation was to better understand coral adaptation by examining these changes. To this end, developments in DNA sequencing technology (Next Generation Sequencing; NGS) provide a powerful observational tool, permitting genome-scale interrogations of genotype, gene expression, and chromatin modifications at low enough cost for ecological applications. I applied this technology to study how variation in genotype, gene expression, and DNA methylation relate to coral fitness under changing environmental regimes. The dissertation is divided into three chapters. The first concerns genetic adaptation to climate change in the Great Barrier Reef. The second characterizes patterns of DNA methylation in a branching coral, and their relation to gene expression plasticity. The third chapter examines the relative importance of all three genomic factors: genotype, gene expression, and DNA methylation, in predicting coral fitness following transplantation to environmentally distinct reefs.

Chapter 1: Genomic determinants of coral heat tolerance across latitudes

ABSTRACT

As global warming continues, reef-building corals could avoid local population declines through "genetic rescue" involving exchange of heat-tolerant genotypes across latitudes, but only if latitudinal variation in thermal tolerance is heritable. Here we show up to ten-fold increase in thermal tolerance of coral larvae when their parents come from a warmer lower-latitude location. We also developed the densest linkage map to date for any coral species using RAD sequencing. We coupled the linkage map with an artificial selection experiment to identify two genomic regions strongly responsive to selection for thermal tolerance in inter-latitudinal reciprocal crosses. These results demonstrate that variation in coral thermal tolerance across latitudes has a strong genetic basis and could serve as raw material for natural selection.

INTRODUCTION

Worldwide, coral reefs are threatened by increasing temperatures associated with climate change (Hughes et al. 2003; Hoegh-Guldberg et al. 2007). Models predict that even a modest increase in the thermal tolerance of reef-building corals over 40-80 years would lower their extinction risk dramatically (Logan et al. 2014). Corals are capable of physiological acclimatization to elevated temperature, and it has been argued that in such long-lived organisms acclimatization rather than genetic adaptation will play the leading role in their response to climate change (Palumbi et al. 2014). Here we present data for the heritable basis of temperature tolerance that supports the potential for rapid adaptation at the genetic level based on standing genetic variation.

Many coral species maintain high genetic connectivity across thousands of kilometers and inhabit latitudinal ranges that span considerable temperature gradients (Ayre and Hughes 2000; Van Oppen et al. 2011). However, it remains unclear to which extent latitudinal variation in coral thermal physiology is heritable and could fuel genetic rescue via exchange of temperature tolerant

immigrants across latitudes (Ingvarsson 2001). Here we use quantitative genetic and quantitative trait loci analyses to address this question in *Acropora millepora* corals from thermally divergent locations separated by five degrees of latitude: Princess Charlotte Bay (PCB) and Orpheus Island (OI, Figure 1A).

Ten crosses were established according to a diallel scheme by cross-fertilizing gametes from four adult colonies from the two locations (Figure 1B). Larval families were cultured in triplicate for 5 days until embryonic development was complete and sampled for tag-based RNA-seq analysis (Meyer et al. 2011). Separately, larval crosses were scored for heat tolerance, measured as odds of survival over 27-31h at 35.5ºC. The target temperature was reached by ramping over 12 hours at the rate of 0.63ºC per hour, less than half of the warming rate on a reef flat during a tidal cycle (Palumbi et al. 2014).

METHODS

Environmental data

The temperature data for the reef flats of Princess Charlotte Bay location (Wilkie Island) and Orpheus Island (Cattle Bay) were obtained from AIMS temperature logger data archive (www.aims.gov.au). For Cattle Bay the logger was placed in 3.5 m water and monthly averages were based on 17 years of data (1993 – 2010). For Wilkie Island the logger was found in 3.7 m of water and averages were based on 11 years on data (2000 – 2012).

Larval rearing

The coral larval cultures were established and reared as described previously (Meyer et al. 2009), with one modification of using less dense cultures (0.5 larvae per mL). Ten crosses were established and reared in triplicate cultures. The two crosses (DA and DB) that are missing from the complete 4x4 diallel design scheme (Fig. 1 B) were not possible to establish due to the limited

amount of eggs released by the colony D. At midday on the 5th day 50-70 larvae from each culture (n=30 larval samples) were preserved in 96% ethanol for tag-based RNA-seq.

Scoring larvae for heat tolerance

On the 6th day post fertilization, samples of 20 larvae (6 replicate samples per culture) were placed in netted well plates allowing water and waste exchange with the large volume (60L) of filtered seawater in the water bath. The temperature was ramped from 28 to 35.5^{o}C over 12 hours and the surviving larvae were counted every 4-6 hours over the next 37 hours. No larval mortality was observed in the control experiment set at ambient temperature (28^{o}C). One of the 30 cultures (DA2) was accidentally lost on the 6th day and was not scored for heat tolerance.

Statistical analysis of larval survival data

The larval counts were analyzed at 27 and 31-hour time points, when the differences in survival between larval families were the most pronounced (Fig. 2A). To assess the effect of PCB parentage on heat tolerance, an over-dispersed binomial mixed model with fixed factors timepoint (27 or 32 hours, used as categorical factor levels), and pcb.parent ("none", "dam", "sire", or "both") was fitted to the counts data using the package MCMCglmm in R (Hadfield 2010). The model also included a scalar random effect of culture to account for repeated measures across timepoints. The p-values for pairwise comparisons between pcb.parent factor levels were calculated based on samples from posterior distributions of parameter values. To evaluate the proportion of variance explained by genetic factors, another over-dispersed binomial mixed model was constructed in MCMCglmm, with a single fixed effect of timepoint (27 or 31), and sire, dam and their interaction as random effects; plus a scalar random effect of culture to account for repeated measures. The model used identical weakly informative inverse Wishart priors (V =1, nu=0.1) for all variance components to indicate that some variance was expected for each. The proportions of variance attributable to parental effects were calculated based on the samples from posterior distributions of the corresponding variance components.

Heat selection experiment

Aliquots of two replicate cultures of each of the two reciprocal crosses (CA and AC), containing about 1,000 larvae in 2L of filtered seawater, were subjected to the same temperature treatment as during larval heat tolerance profiling (35.5^{o}C, ramped from 28^{o}C over 12 hours). The treatment continued for approximately 36 hours until 30-50 surviving larvae remained in each of the cultures, which were collected for individual genotyping. As a control, 50 larvae were sampled from portions of the same larval cultures that did not undergo heat treatment. The resulting samples comprised 8 separate groups corresponding to 2 reciprocal crosses, 2 larval cultures per cross, and 2 treatments (heat-selected and control) for each culture, 30-50 larvae per treatment.

2bRAD genotyping

2bRAD is a whole-genome genotyping methodology that relies on restriction endonucleases to target the genotyping effort to a small fraction (about 1-5%) of the genome (Wang et al. 2012). The larval and adult coral DNA was isolated using a phenol-chloroform protocol (Davies et al. 2013). A BcgI restriction enzyme was used to generate fragments for sequencing. Sequencing of RAD fragments was performed on an Illumina HiSeq 2500 instrument and generated a total of 224,591,655 raw sequence reads. Reads were 'demultiplexed' based on barcodes incorporated during the ligation and PCR amplification stages. Dual adapters used for demultiplexing reads from offspring were trimmed using a custom perl script. All reads were quality filtered using Fastx Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/). After trimming and quality filtering a total of 180,957,303 reads had been generated from 361 libraries, with an mean count of 501,267 reads ± 39,502 SE. 23 offspring libraries had read counts < 75,000 and were removed from the analysis. The two parental individuals along with two additional adult samples were sequenced in triplicate and with greater read depth. Mean read counts for these 12 samples was 3,759,904 ± 478,584 SE. Mean depth for the offspring samples was 413,525 ± 18,895 SE. Of the 326 offspring samples included in the analysis, 190 were control samples and 136 were heat

selected. The detailed wet lab protocol, bioinformatic pipeline and custom scripts for performing 2bRAD analysis are provided as supplementary file 6. The current version of the methodology, including a novel modification allowing for removal of PCR duplicates, can be downloaded from https://sourceforge.net/projects/gatk-2brad/ .

Variant calling

Variant discovery, genotyping and filtering was performed using The Genome Analysis Tookit version 3.1-1 (GATK) (McKenna et al. 2010). The basic pipeline framework includes mapping, realigning the reads around indels, recalibrating base quality scores, calling putative variants, then developing an adaptive filtering model to select highly confident variants (DePristo et al. 2011). Descriptions of the tools, usage, and best practices can be found on the GATK website (https://www.broadinstitute.org/gatk/). The following is a description of our adaptation of this framework.

Reads were mapped using Bowtie2 (Langmead and Salzberg 2012) with the –local option against a draft assembly of the *A. millepora* genome released to the public by David Miller and coworkers (James Cook University) in July 2011 under an 18-month embargo. The alignment files generated from Bowtie2 were used as input for the GATK variant calling pipeline. Following realignment around indels, a first round of putative variants was generated using GATK's UnifiedGenotyper. From these initial variants calls, the top 10^{th} percentile for quality was selected using a custom python script. These high confidence base calls were passed as the 'knownSites' to GATK's BaseRecalibrator, which recalibrates base quality scores to remove biases of particular sequencing technologies and contexts (DePristo et al. 2011). The realigned and recalibrated reads were then used to generate a second set of putative variants.

The next step is the variant recalibration step, which provides estimates of the probability that each putative variant is an actual SNP. These estimates are used to filter the putative variants to minimize false discoveries. The VariantRecalibrator program generates an adaptive error model

based on a set of user supplied set of "true" SNPs and a chosen set of VCF annotations such as read depth or inbreeding coefficient (DePristo et al. 2011). The source of the "true" SNPs supplied to the VariantRecalibrator should be one that is well established for the particular organism. Lacking an established set of SNPs for *A. millepora*, we used the variants that were reproducibly genotyped across the technical replicates of our four parental colonies instead (3 replicates x 4 colonies = 12 genotyping sets). For a locus to be considered a "true variant" for the recalibration procedure, out of 12 genotyping sets not more than three (25%) could be missing data for the locus, and all the non-missing calls should agree among replicates. The result was 9,338 variant sites (out of 68,141) that were both polymorphic among the adults and had fully consistent genotype calls between technical replicates. This set was used as the 'true' sites for the variant recalibration. The statistics used to build the error model in our case were total unfiltered depth of coverage (DP), inbreeding coefficient (InbreedingCoeff), quality by depth (QD), and the root mean square of the mapping quality of the reads across all samples (MQ). The filtering process represents a compromise between maximizing sensitivity for recognizing the "true" variants and minimizing false discoveries. To facilitate filtering, the VariantRecalibrator generates tranches of the variant calls with varying truth sensitivities (the percentage of "true" variants captured in the tranche). The false discovery rate for each tranche is estimated based on the difference between an expected transition/transversion (Ti/Tv) ratio and the Ti/Tv ratio for the variants included in the tranche. Tranches with Ti/Tv ratio lower than the expected value are inferred to include false positives. The target Ti/Tv value in *A. millepora* was 1.41, estimated from the same 9,338 reproducibly genotyped SNPs that were used to build the error model. In applying the recalibration we chose the tranche with 96% truth sensitivity, which had a Ti/Tv ratio of ~1.43. The final output of the genotyping pipeline was a set of 55,211 filter passing variants recorded in VCF format.

Selection of variants for mapping

The following is a description of additional filters used to prepare the genotype data for linkage mapping. These filters were applied using custom python scripts and are described in order

of their implementation. Using vcftools (Danecek et al. 2011) individual genotype calls with a sample depth of less than 6 reads were changed to missing data, to reduce the incidence of false homozygotes. Assuming equal probabilities of sequencing the paternal and maternal chromosome at a given site, the probability of not encountering a second allele (potentially leading to a false homozygote call) can be described as $2^{-(n-1)}$, where n is the depth of coverage for the individual at that site. We chose a depth threshold of six, with an estimated probability of missing a second allele of 1/32 as a comfortable compromise between excluding too many genotype calls and having too little confidence in homozygote calls. We next filtered for variant quality based on agreement between technical replicates. Variants that had missing data for two of the three replicates for either parent were removed (27859 variants). Variants with more than one disagreement between parent replicates were removed (544 variants). In cases of a single disagreement between replicates the most common genotype was retained as the consensus type for that parent. Variants for which the consensus genotypes for both parents were homozygous were then removed because they were not informative for linkage analysis using the outbred F1 design (13735 variants). Next we filtered the remaining genotype calls that were incompatible with the parental genotypes. First genotypes with alleles not present in the parental genotypes for that locus (1223 total) were assumed to be sequencing errors and changed to missing data. In addition, a small proportion (2.8%) of the filter-passing genotypes calls were incompatible based on their combinations of parental alleles. A large proportion of these (86%, or 2.4% of all calls) could be inferred as false homozygotes. For instance, in a cross represented by AAxAT, an offspring genotype call of TT should not occur, and is strongly indicative of a false homozygote from a failure to sequence the second allele. Such genotypes were changed to the inferred heterozygote (in the example above, TT genotype call would be changed to AT). The remaining 0.4% of incompatible genotypes could not be resolved and were changed to missing data. Variants genotyped in less than 100 offspring were removed (8174 variants). Filtering for deviation from Mendelian expectations was performed using chi square tests with a threshold of $P = 0.05$. Because we wanted our map to include markers displaying segregation distortion due to heat selection, the Mendelian filtering was performed

based only the control samples. As a result 1854 variants showing significant distortion among the controls were excluded, while variants showing distortion specifically among the heat selected samples were retained. The input for linkage analysis contained a total of 3047 passing variants.

Linkage mapping

Linkage analysis was carried out using JoinMap 4.1 (Stam 1993). Markers were categorized by cross types using the annotations given for the outbreeder full sib family protocol in the JoinMap user manual (Ooijen 2006). Offspring from both reciprocal crosses were used for the analysis and all markers were coded as if parent A was the female and parent C was the male. Initial grouping was carried out using the independence LOD score. A grouping threshold of 12 produced 14 linkage groups corresponding to the haploid number of chromosomes for *A. millepora* (Kenyon 1997). The independence LOD is based on the G^2 statistic for independence and is not affected by segregation distortion (Ooijen 2006). After initial groups were established, ungrouped markers were incrementally added to their best fitting linkage groups by walking from an LOD threshold of 11 down to 6.

Ordering of loci was performed using the regression mapping algorithm. This algorithm orders the loci by adding them one at a time to their best fitting positions starting with the most informative markers (Stam 1993). To avoid locally optimal orders, the ripple function was applied after adding each locus. The Kosambi mapping function was used to convert the recombination frequencies into map distances (centiMorgans).

Genomic regions responding to heat selection

By treating a subset of our larval cultures with heat we intended to select for alleles that conferred increased larval tolerance to critically elevated temperatures. If such selection occurred, we would expect to see 1) reproducible differences in allele frequencies between heat selected and control cultures and 2) spatial autocorrelation between RAD loci and the magnitude of selection signal. The second prediction should result from linkage between RAD loci and any alleles that were

under selection due to heat treatment. For the first prediction, we first used Fisher's exact tests (Fisher 1922) to compare allele counts between each replicate pair of control and heat treated cultures. *P*-values from replicates were then combined using Fisher's Combined Probability method (Fisher 1932). When using this method, it is important to use one-tailed *P*-values so that information on the direction of effect is retained (Whitlock 2005). This required that we use one-tailed tests for comparisons of allele frequency. To obtain the predicted direction of selection (either for or against the reference allele) at each locus, we calculated the average difference in reference allele frequency ($p_{\text{heat selected}} - p_{\text{control}}$) for the two replicates for each cross. The sign of the average was used to designate the alternative hypothesis for one-tailed Fisher's exact tests. To compensate for this *post hoc* use of the data to inform one-tailed tests, we multiplied the *P*-values by 2. Combined *P*-values were adjusted to control for false discovery rate (Benjamini and Hochberg 1995). The genotype calls used to perform all selection analyses were generated using the same filtering parameters described above, with the exception that a read depth threshold of four rather than six was used to change genotypes to missing data.

To visualize genomic regions showing signatures of selection Manhattan plots were generated for each genetic cross by –log transforming the combined *P*-values and plotting them against their predicted chromosomal locations. To test for spatial autocorrelation of selection signal we used a bootstrap approach. First a null distribution was generated from 10^5 permutations of randomly selecting 15 markers (without replacement) from the dataset and counting the number with Fisher's combined *P*-values less than 0.05. Then each linkage group was divided into equal non-overlapping windows each comprising 15 adjacent RAD loci. For each window, the number of loci with combined *P*-values below 0.05 was counted and compared with the null distribution to provide a bootstrap *P*-value for that window. Following false discovery rate adjustment, window *P*-values less than 0.1 and 0.05 were indicated on the Manhattan plot for each cross. This analysis served to identify additional genomic regions with moderate, but strongly autocorrelated deviations in allele frequency.

Strength of selection

Since in nearly all cases the parental genotypes were of the type AA:Aa, the selection process in the F1 can be viewed as a result of competition between AA homozygote and Aa heterozygote. In such case we calculated selection coefficients using the following equation (Otto and Day 2007):

$$s = \frac{p - p'}{p(1 - p')}$$

where *s* is the decline in fitness of the selected allele relative to the other allele, *p* is the frequency of the negatively selected genotype in the absence of heat selection, and *p'* is the frequency of the genotype after heat selection.

Genes under peaks of the Manhattan plot

Gene coordinates were inferred using the reference transcriptome (Moya et al. 2012) and the draft genome assembly for *A. millepora*. Each contig from the reference transcriptome was aligned to the genome assembly using blastn (Altschul et al. 1997) with setting to only return alignments against the top hit. Alignment coordiantes for each contig were recorded as the minimum and maximum alignment positions. When calculating physical distances between genes and RAD loci we used the midpoint of contig's alignment coordinates. We report genes found within 100 kb to either side of the top-scoring RAD locus within the boundaries of the peak.

RESULTS AND DISCUSSION

Survival rates varied substantially among families (Figure 1D). A mixed effects generalized linear model with random effects of sire, dam and their interaction as predictors indicated that the combined parental effects (i.e., broad-sense heritability) accounted for 87% of total deviance in odds of larval survival (Fig. 1 E, 95% credible interval of the posterior: 75-99%). Proportions of deviance due to sire, dam and their interaction were estimated at 11%, 66% and 12%, respectively, although the credible intervals were wide due to the limited scope of our crossing design (Fig. 1 E). Importantly, parents from the warmer location (PCB) conferred

significantly higher thermo-tolerance to their offspring relative to parents from the cooler location (OI), with a PCB dam conferring a 5-fold increase (P_{MCMC}<0.001) and a PCB sire conferring an additional 2-fold increase (P_{MCMC}=0.048) in survival odds (Figure 1F).

To further demonstrate that larval heat tolerance has genetic basis and can respond to selection, we quantified genomic effects of artificial selection by heat in two inter-latitudinal reciprocal crosses (AC and CA). Selected samples consisted of the last 30-50 heat stress surviving larvae out of the initial ~1000, while control samples consisted of 50 larvae from unstressed cultures. This experiment was performed with two culture replicates from each cross, resulting in eight compared groups. Larvae were individually genotyped (n = 326) using 2bRAD methodology (Wang et al. 2012) to construct a genetic linkage map and identify genomic regions displaying reproducible allele frequency shifts in response to heat selection.

The linkage map contained 1448 markers in 14 linkage groups, and had the total length of 1358 centiMorgans (cM) (Figure 2). In both crosses, the selection was predominantly against paternally-derived haplotypes (Figure 3), resulting in markedly different genome-wide patterns of selection between reciprocal crosses (Figure 2A-B). The strength of negative selection, measured as a decrease in survival of larvae bearing the less preferred haplotype, reached unity in LG 10 in the CA cross (i.e., the less preferred haplotype was completely eliminated from the larval pool) and 0.91 in LG 5 in the AC cross. No statistically significant signatures of selection were observed when comparing pairs of unselected samples (Figure 2C). Selection against paternal haplotypes aligns well with the putative involvement of mitochondria in heat tolerance determination: such selection could be due to poor compatibility of certain paternal nuclear alleles with maternal mitochondria under stress (Hoekstra et al. 2013).

Our study demonstrates, for the first time, heritability of coral stress-related phenotypic and molecular traits and thus highlights the adaptive potential stemming from standing genetic variation in coral meta-populations. Two lines of evidence point towards the importance of mitochondria and mitochondrial-nuclear interactions in determining heat tolerance, including its

predominantly maternal inheritance (Figure 1E), persistent selection against paternal haplotypes in reciprocal crosses under heat stress (Figure 2 and Figure 3). High maternal effect on larval thermal tolerance could also be partially due to epigenetic modification, which remains poorly understood in corals. Most importantly, the strong response of two genomic regions to heat selection (Figure 2) directly confirms that natural variation in heat tolerance is both heritable and evolvable. The genetic rescue scenario, therefore, emerges as a plausible mechanism of rapid coral adaptation to climate change, especially if the natural connectivity of corals across latitudes is enhanced by assisted colonization efforts (Hoegh-Guldberg et al. 2008).

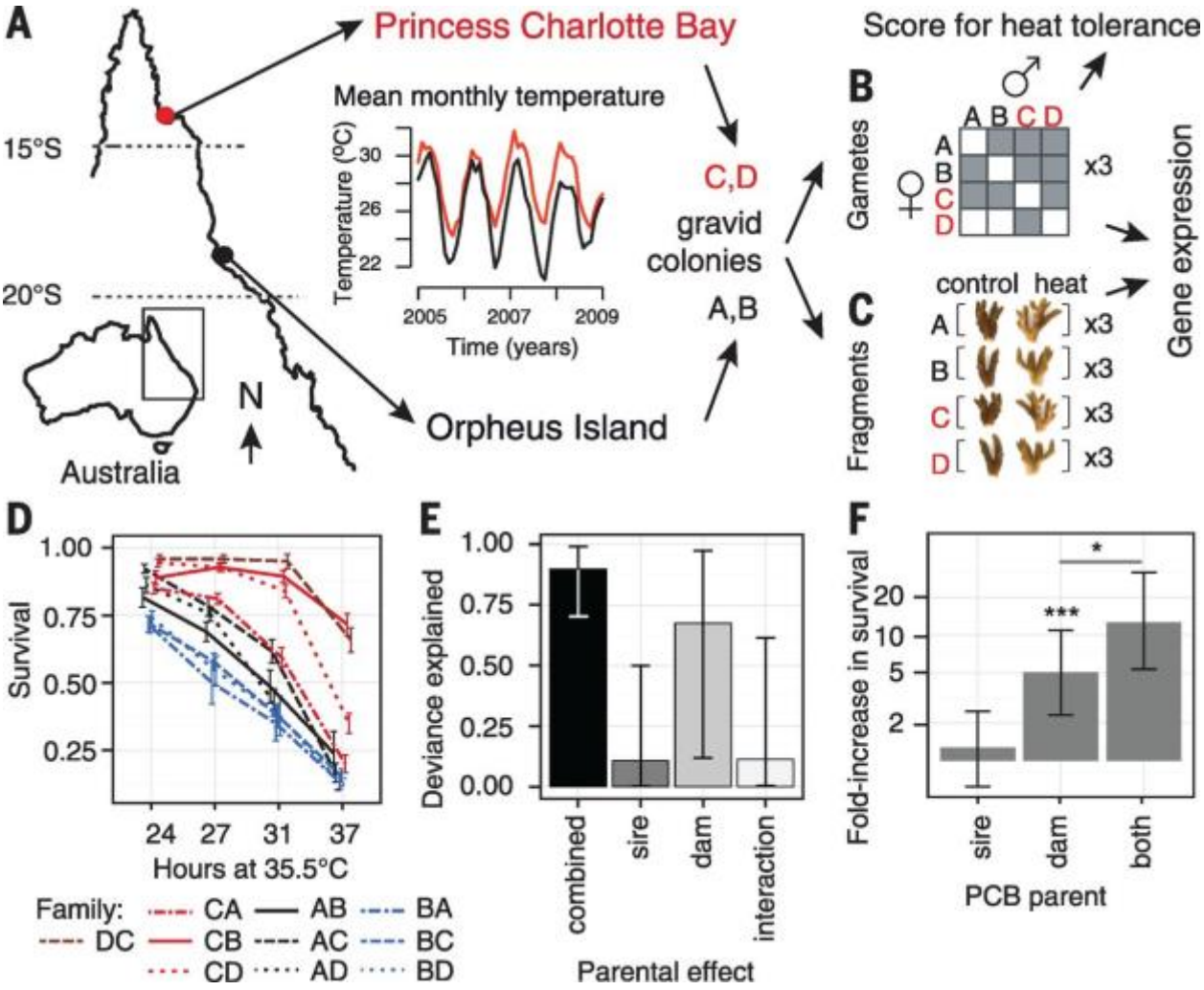

Figure 1 Experimental design and quantitative genetics of larval heat tolerance. (A) Sampling locations and their annual temperature regimes on the Great Barrier Reef, Australia. (B) Crossing design matrix where solid squares represent established crosses. (C) Experimental design to quantify gene expression differences between parental colonies under heat stress (31.5°C for 3 days). (D) Mortality curves ± SE for each larval family. In the family identifier, the first letter is dam (mother); the second letter is sire (father). (E) Proportion of total deviance explained by parental effects. (F) Increase in odds of larval survival with parents from the warmer location (PCB) relative to the larvae with both parents from the cooler location (OI). ***$P < 0.001$, *$P < 0.05$. Whiskers on (E) and (F) denote 95% credible interval of the posterior.

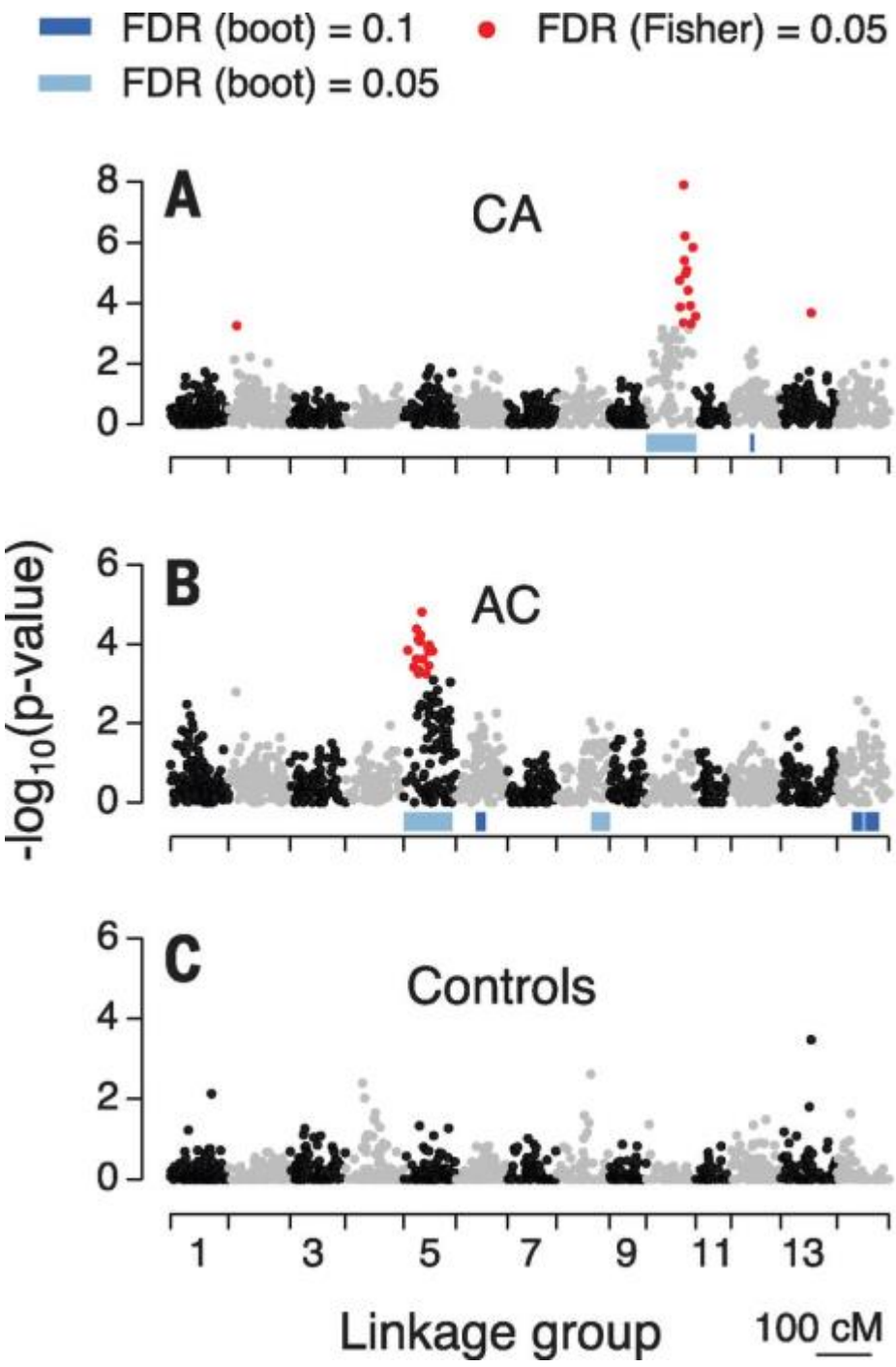

Figure 2 Manhattan plot of allele frequency difference after selection by heat. (A) Selection effects in CA family. (B) Selection effects in AC family. (C) Differences in allele frequencies among control samples. Red points show markers at 5% FDR according to the Fisher's combined probability test; blue bars identify regions with significant clustering of such markers (according to 100,000 bootstrapped replicates).

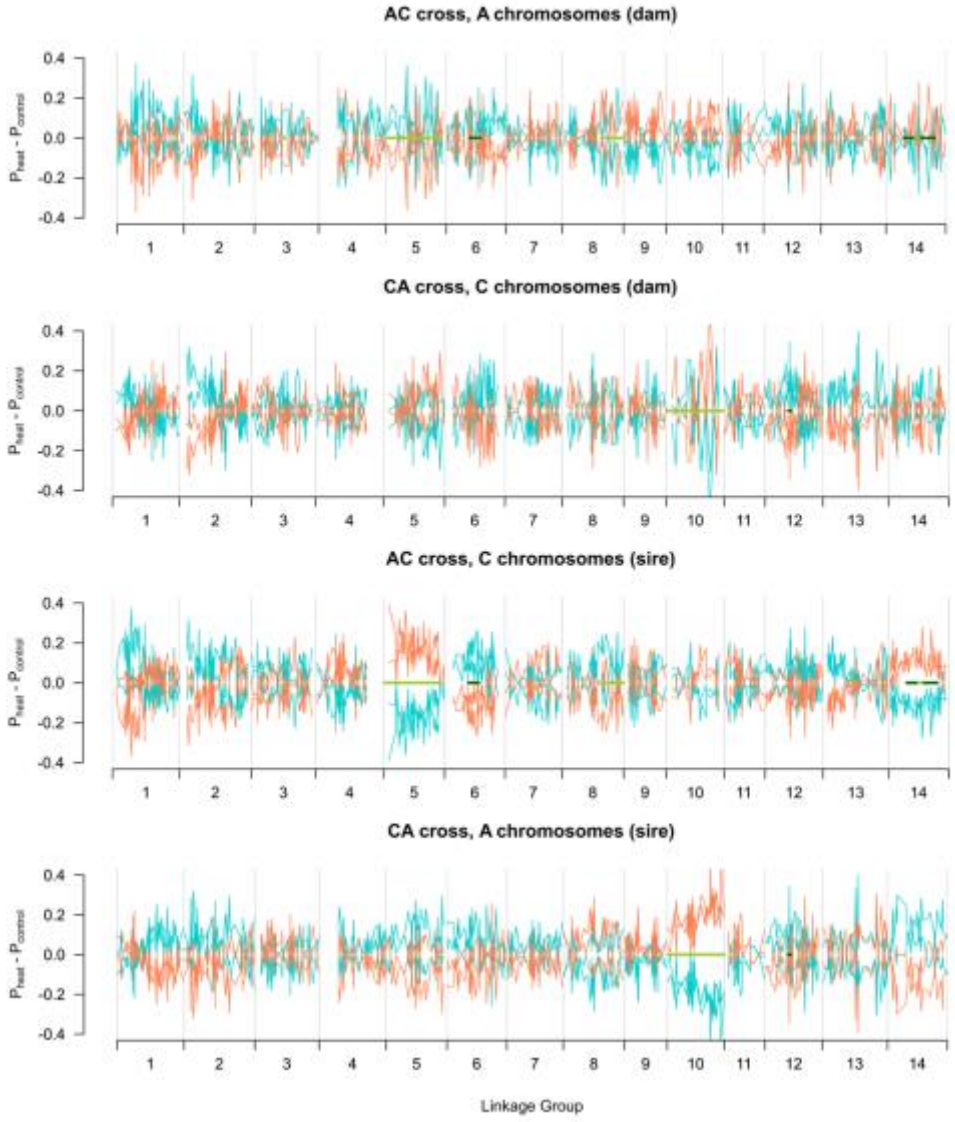

Figure 3 Allele frequency changes in larval cultures as a result of heat selection. In each panel, the X-axis gives position of markers in the linkage map and the Y-axis is allele frequency change in selected cultures compared to unselected controls. Panels (A) and (C) show the change in frequencies of maternally-derived variants (i.e., SNPs that were heterozygous in the dam), panels (B) and (D) - the change in frequencies of paternally-derived variants. The two lines of the same color on each panel represent two replicates of the heat stress experiment within each cross; red and blue lines correspond to different haplotypes in the dam (A, C) or sire (B, D). The haplotype color is consistent among the crosses, that is, a blue paternal haplotype in the AC cross is the same as the blue maternal haplotype in the CA cross. The green bars identify regions that show bootstrap-supported significant clustering of low p-values (obtained by Fisher's exact test for each replicate followed by Fisher's combined probability test; light green – 5% FDR, dark green – 10% FDR).

PREFACE TO CHAPTER 2

This chapter is divided into two sections to reflect the chronology and methodology of the research it describes. The first section describes an exploratory study in which patterns of DNA methylation were not assayed directly, but inferred from sequence data. The second section describes a follow-up study that confirmed results from the exploratory study and explored additional evolutionary consequences of DNA methylation in the coral genome.

Chapter 2a: Using an evolutionary signature to characterize of gene body methylation in *Acropora millepora*

ABSTRACT

In invertebrates, genes belonging to dynamically regulated functional categories are often less methylated than "housekeeping" genes, suggesting that DNA methylation may modulate gene expression plasticity. To date, however, experimental evidence to support this hypothesis across different natural habitats has been lacking. Gene expression profiles were generated from 30 pairs of genetically identical fragments of coral *Acropora millepora* reciprocally transplanted between distinct natural habitats for 3 months. Gene expression was analyzed in the context of normalized CpG content, a well-established signature of historical germline DNA methylation. Genes with weak methylation signatures were more likely to demonstrate differential expression based on both transplantation and population of origin than genes with strong methylation signatures. Moreover, the magnitude of expression differences due to environment and population were greater for genes with weak methylation signatures. Our results support a connection between differential germline methylation and gene expression flexibility across environments and populations. Studies of phylogenetically basal invertebrates such as corals will further elucidate the fundamental functional aspects of gene body methylation in Metazoa.

INTRODUCTION

Phenotypic plasticity refers to the ability of an individual to adjust its phenotype in response to environmental cues (Kelly et al. 2012). Under changing environmental conditions, theory predicts that phenotypic plasticity may mitigate loss of fitness (Chevin et al. 2013), and facilitate evolutionary adaptation (Price et al. 2003; Yeh and Price 2004). For sessile organisms such as plants and corals, plasticity is predicted to be of particular importance, as these organisms cannot migrate away from suboptimal environments (Nicotra et al. 2010; Barshis et al. 2013b). In the table top coral *Acropora hyacinthus*, colony fragments with previous exposure to elevated

temperatures demonstrate increased bleaching resistance, suggesting an important role of plasticity in coral heat tolerance (Palumbi et al. 2014). Hence predicting the future of reef-building corals and the ecosystems they support requires an understanding of their mechanisms of phenotypic plasticity. In most marine organisms, however, the molecular mechanisms that translate environmental stimuli into appropriate cellular responses are poorly understood (Aubin-Horth and Renn 2009). One possible plasticity-modulating mechanism that is yet to be investigated in corals is DNA methylation (Angers et al. 2010; Roberts and Gavery 2012).

DNA methylation is a widely conserved epigenetic modification involved in eukaryotic gene regulation (Richards 2008). In mammals, the majority (70-80%) of CG dinucleotides are methylated, with the exception of stretches of sequence rich in CG dinucleotides called CpG islands (CGIs) (Tucker 2001). CGIs are generally not methylated, but can be targeted for methylation under particular conditions (Jaenisch and Bird 2003). When methylation of CGIs occurs, its effect on transcription is based on the proximity of the CGI to a transcription start site (TSS), inhibiting initiation of transcription when near the TSS but not when far away from it (Jones 2012). There is evidence that this form of epigenetic regulation is involved in genome-environment interactions. DNA methylation is associated with persistent stress-induced gene expression in mice (Murgatroyd et al. 2009) and humans (Heim and Binder 2012), as well as in plants (Wang et al. 2011). It has also been linked with variation in disease development (Portela and Esteller 2010) and mediating phenotypic differences between monozygotic twins (Javierre et al. 2010; Miyake et al. 2013).

In contrast to nearly ubiquitous methylation in mammalian genomes, genomic methylation in many invertebrates occurs specifically on CpG dinucleotides within gene bodies (also called transcription units) (Suzuki et al. 2007; Zemach et al. 2010). Within gene bodies, methylation occurs primarily on exons rather than introns (Lyko et al. 2011; Bonasio et al. 2012). Density of gene body methylation is not equivalent across genes. Studies of multiple invertebrate taxa report bimodal patterns of gene body methylation, in which genes are separated into hypermethylated

and hypomethylated classes (Zemach et al. 2010; Falckenhayn et al. 2013; Sarda et al. 2012). Analyses of gene ontology (GO) terms in the context of these methylation classes have demonstrated characteristic divisions based on gene function. Basic biological functions with similar regulatory dynamics across tissue types and developmental stages tend toward strong methylation. Examples of basic biological functions include cellular metabolic processes, nucleic acid metabolism, and translation (Sarda et al. 2012; Park et al. 2011). In contrast, functions that are dynamically regulated across tissues and developmental stages tend toward sparse methylation (Elango et al. 2009; Hunt and Brisson 2010). Examples of dynamically regulated functions include development, cell-cell signaling, and signal transduction (Gavery and Roberts 2010; Park et al. 2011). These findings suggest that bimodal gene body methylation may regulate flexibility of gene expression, with strongly methylated genes marked for stability and weakly methylated genes marked for flexibility. DNA methylation has also been linked with phenotypic plasticity, most strikingly in caste development in honeybees *Apis mellifera*, which is dependent on larval diet and activity of *de novo* DNA methyl-transferase (DNMT3) (Kucharski et al. 2008; Foret et al. 2012). Caste-specific genes in honeybees (i.e. genes with significant differential expression between queens and workers) are significantly biased toward weak methylation (Elango et al. 2009).

These findings have led to the hypothesis that invertebrate DNA methylation is involved in regulating environmentally driven gene expression and phenotypic plasticity (Angers et al. 2010; Roberts and Gavery 2012). Among marine invertebrates, this hypothesis has yet to be validated in natural ecological contexts. In this study, we predicted that environmentally flexible gene expression in the branching coral *Acropora millepora* would be associated with signatures of weak gene body methylation. To test this prediction we analyzed gene expression profiles from clonal colony fragments reciprocally transplanted between distinct natural habitats. We show that elevated CpG content, a signature of historically weak germline methylation, is linked with environmentally driven gene expression. Our results suggest a potential role of DNA methylation

in modulating the balance between stable gene expression required for homeostasis and flexible expression required for plasticity.

METHODS

Genomic resources

Coding sequences were extracted from the *A. millepora* transcriptome (Moya et al. 2012) using the script CDS_extractor.pl which is part of the Matz lab's transcriptome annotation bundle available on the Matz lab website at http://www.bio.utexas.edu/research/matz_lab/matzlab/Methods_files/transcriptomeAssemblyAnnotation.v.0.5.tgz and on GitHub (doi:10.5281/zenodo.12232). This script uses Blastx (Altschul et al. 1997) results against a protein sequences database to identify open reading frames and extract them while correcting for occasional frame shifts. For the protein reference database we combined the proteomes of *Nematostella vectensis* (Putnam et al. 2007) and *Acropora digitifera* (Dunlap et al. 2013). Blastx against this database was performed with an evalue cutoff of 1e-4. Based on manual verifications of a subset of *A. digitifera* protein annotations, they were pre-filtered to include sequences longer than 60 amino acids with the annotation assigned based on the listed e-value = 1e-20 or better. Annotation of our coding sequences with GO terms was based on Blastx hits to the annotated *Nematostella vectensis* (Putnam et al. 2007) and *Acropora digitifera* (Dunlap et al. 2013) proteomes as part of the annotation pipeline indicated above. All GO annotations associated with the best hit from these references were transferred to the query sequence in our dataset. GO annotations were further supplemented with GO terms based on BLASTx matches (e-value cutoff of 10^{-10}) to the SwissProt Protein database (Consortium 2013). Coding sequences with no annotations were excluded from the analysis. As gene length can influence the validity of $CpG_{o/e}$ as a predictor of methylation status (Wang et al. 2013), and DNA methylation tends to decrease toward the 3' end of invertebrate gene bodies (Zemach et al. 2010; Bonasio et al. 2012), we controlled for differences in transcript length by examining CpG content only in the first 1 kb

of the coding regions of each gene. Sequences shorter than 300 bp (2500 in total) were also excluded from the analysis. CpG values were bounded between 0.001 and 2, so that 1 gene with a value of 2.08 was excluded and 89 with values below 0.001 were excluded.

CpGo/e class assignment

To predict gene body methylation in *A. millepora*, we used normalized CpG content ($CpG_{o/e}$) calculated as in Elango *et al*. (Elango et al. 2009).

$$CpG_{o/e} = \frac{P_{CpG}}{(P_C \times P_G)}$$

Where P_{CpG}, P_C and P_G are the frequencies of CpG dinucleotides, C nucleotides, and G nucleotides respectively. Because 5-methylcytosines are hypermutable (Sved and Bird 1990), and invertebrate DNA methylation is targeted specifically to CpG dinucleotides, sequences that are methylated in the germline are predicted to become CpG deficient over evolutionary time (Flores and Amdam 2011). As a result $CpG_{o/e}$ is inversely related to the degree of historical germline methylation. Low $CpG_{o/e}$ values indicate strong methylation while high $CpG_{o/e}$ values indicate weak methylation. This metric has been shown to correlate with direct measures of DNA methylation in a number of animal models (Zemach et al. 2010; Wang et al. 2013; Sarda et al. 2012; Xiang et al. 2010).

To test whether the distribution of $CpG_{o/e}$ values was best described as mixture of distributions we used the package Mclust (Fraley and Raftery 2007) implemented in the R environment (R Core Team 2015). Bayesian information criterion (BIC) was used to compare the likelihood of Gaussian Mixture Models with different numbers of components. We found that a two-component model provided substantially better fit than a single component model [see additional file 1 for trace of BIC for different numbers of components]. The R package Mixtools (Benaglia et al. 2007) was used to fit and trace the two-component mixture model of the distribution. Following Hunt et al. (Hunt and Brisson 2010), we used the point of intersection of the two component curves to separate genes into two components: the 'low-CpG component'

(predicted to be hypermethylated) and the 'high-CpG component' (predicted to be hypomethylated).

Analysis of CpGo/e for biological processes

To analyze patterns of gene function relative to predicted methylation, we assigned annotated genes to 14 general biological processes based on gene ontology (GO) terms. The GO terms were selected to match a subset of those analyzed by Gavery and Roberts (Gavery and Roberts 2010). To these we added four additional hand-picked GO terms: *ribosome biogenesis*, *translation*, *defense response*, and *regulation of response to stimulus*, to better demonstrate the spread of "house-keeping" versus dynamic biological processes. A single gene could be assigned to multiple GO terms. GO terms with fewer than 20 representative genes, including *death* (no annotated genes) and *protein metabolism* (11 genes), were excluded from the analysis. Nonrandom sorting of genes from each function between the high-$CpG_{o/e}$ and low-$CpG_{o/e}$ components was ascertained using Fisher's exact test (Fisher 1922).

Reciprocal transplantation experiment

To test for a relationship between environmentally induced gene expression and predicted methylation, we analyzed gene expression patterns of samples from a reciprocal transplantation experiment. Briefly, reciprocal transplantations were made between two environmentally distinct study sites (Keppel: 23˚09S 150˚54E and Orpheus 18˚37S 146˚29E) separated by 4.5 degrees of latitude in the Great Barrier Reef. On the 23rd April (Orpheus) and 4th May (Keppels) 2010 fifteen colonies were collected from wild populations from each site and divided into two. One half of each colony was replaced in its native habitat, while the second half was transplanted to the alternate study site. Samples from all coral fragments were collected at midday after three months (9th July 2010 at Orpheus, 14th July at Keppels) frozen in liquid nitrogen, then transferred into RNAlater (Ambion, Austin, TX, USA) for gene expression profiling.

Analysis of gene expression

The gene expression in 56 samples (four samples failed to produce gene expression profiles and were not included) from the reciprocal transplant experiment was profiled using tag-based RNA-seq library method (Meyer et al. 2011). The latest version of the protocol and bioinformatics pipeline are available at http://sourceforge.net/projects/tag-based-rnaseq/. Briefly, the method works by sequencing short fragments from the 3' end of mature mRNA transcripts by priming off the poly-A tail during first stand cDNA synthesis. Transcript abundance is inferred through normalized fold coverage of reads mapped to the species transcriptome (Moya et al. 2012). While the method is exceptionally cost effective it does not allow analysis of relative expression of different transcript isoforms. Sequencing was performed using the SOLiD system producing a total of 108499924 filter passing reads (average reads per sample = 1937499 ± 788474 standard deviation). The counts data were analyzed using generalized linear model implemented in DESeq package (Anders and Huber 2010) with the factors "origin" and "transplant location". The R script used to implement DESeq is available at [DOI](https://zenodo.org/badge/doi/10.5281/zenodo.12626.png)](http://dx.doi.org/10.5281/zenodo.12626) and on GitHub at https://github.com/grovesdixon/CpGoe.git. False discovery rate was controlled at 1% (Benjamini and Hochberg 1995). Enrichment of differentially expressed genes (i.e., showing significant effect of transplant location or origin) in the high-$CpG_{o/e}$ component was tested using Fisher's exact test (Fisher 1922). To assess this association on a continuous scale, we plotted transplantation effect and origin effect estimated for each gene in DEseq against its $CpG_{o/e}$. Effect size for transplantation was estimated as [log(mean expression of samples placed at Orpheus / mean expression samples placed at Keppel)], whereas the origin effect was estimated as [log(mean expression of samples that originated at Orpheus / mean expression samples that originated at Keppel)].

Plotting relationships between expression and CpGo/e

To visualize genome wide trends between differential gene expression and predicted methylation (Figure 7A,D) we plotted data from 25 equal-sized gene quantiles based on CpGo/e. The component curves from (Figure 5A) were overlaid to show the relative densities of the high- and low-CpG components. Vertical scaling of the component curves did not correspond to y-axis labels. To plot mean transcript abundance against CpGo/e (Figure 8), genes were divided into 25 equally sized quantiles based on CpGo/e values. For each gene, expression was averaged across all samples. Mean expression for all genes in each quantile was then plotted against the mean CpGo/e value for the quantile.

RESULTS

Normalized CpG content of coding regions is bimodally distributed in A. millepora

Normalized CpG content ($CpG_{o/e}$; see methods) is a well-established evolutionary signature of DNA methylation (Zemach et al. 2010; Wang et al. 2013; Sarda et al. 2012; Xiang et al. 2010). We used this metric to estimate the strength of methylation of the coding regions of 15,221 genes across the *A. millepora* transcriptome. Because 5-methylcytosines are hypermutable (Sved and Bird 1990), and invertebrate DNA methylation occurs specifically on CpG dinucleotides, sequences that are methylated in the germline are predicted to become CpG deficient over evolutionary time (Flores and Amdam 2011). As a result, $CpG_{o/e}$ values are inversely related to the degree of historical germline methylation. Low $CpG_{o/e}$ values indicate strong methylation while high $CpG_{o/e}$ values indicate weak methylation. Importantly, $CpG_{o/e}$ has been shown to strongly correlate with direct measures of somatic DNA methylation in a number of animal models (Zemach et al. 2010; Wang et al. 2013; Sarda et al. 2012; Xiang et al. 2010).

Using this metric, we found a characteristic bimodal pattern in which one set of genes is predicted to have strong germline methylation and a second is predicted to have weak germline methylation. For the distribution of $CpG_{o/e}$ values, a two-component mixture model provided

substantially better fit than a single component model (Figure 4) and as a consequence, we modeled the distribution with a two-component model (Figure 5A). We refer to the two components of the distribution as the low-$CpG_{o/e}$ and high-$CpG_{o/e}$ components. Means for the fitted component curves were estimated as 0.36 for the low-$CpG_{o/e}$ component and 0.74 for the high-$CpG_{o/e}$ component. Genes were assigned to either component based on the intersection of the fitted component curves at 0.46. Genes within the low-$CpG_{o/e}$ component are predicted to be strongly methylated and genes in the high-$CpG_{o/e}$ component are predicted to be weakly methylated. As a control, we showed that normalized content for GpC dinucleotides (which are not expected to be targeted for methylation) was unimodally distributed (Figure 5B). Also, as expected under the predicted mutational pattern for 5-methylcytosine (substitution for thymine as a result of deamination (Sved and Bird 1990)), normalized TpG content showed an inverse linear relationship with $CpG_{o/e}$ (Figure 5C).

CpGo/e shows characteristic associations with different biological processes

To test for associations between $CpG_{o/e}$ and gene function, we sorted genes among different biological processes based on gene ontology (GO) annotations. Mean $CpG_{o/e}$ varied significantly between biological processes (ANOVA $p << 0.0001$) and most processes were enriched in either the low or high-CpG component (Fisher's exact test, Figure 6). Genes associated with RNA metabolism, translation, ribosome biogenesis, DNA metabolism, cell cycle and proliferation, and cellular organization and biogenesis tended toward low $CpG_{o/e}$ values, indicating strong germline methylation. Genes associated with signal transduction, cell-cell signaling, developmental processes, cell adhesion, defense response and regulation of response to stimulus tended toward high $CpG_{o/e}$, indicating weak germline methylation. Genes associated with stress response showed intermediate mean $CpG_{o/e}$ but were enriched in the high-CpG component ($p < 0.05$).

High CpGo/e is linked with environmentally flexible gene expression

To investigate the relationship between environmentally induced gene expression and $CpG_{o/e}$, we used DEseq (Anders and Huber 2010) to compare gene expression between groups of clonal colony fragments reciprocally transplanted between two environmentally distinct natural habitats. Fifteen colonies were collected from each site and divided into two halves. One half of each colony was replaced in its native habitat, while the second half was transplanted to the alternate site (transplant location). In this way each genotype, represented by the two halves of the original colony, was exposed to both sites for a three-month study period. Following three months, global gene expression profiles were generated from all colony halves using tag-based RNA-seq method (Meyer et al. 2011). To detect effects of environment we compared gene expression between samples halves grouped by the transplantation site (the site they were placed at during the study period). In this way, the two groups being compared represented the same 30 genotypes exposed to two different environments for the three-month period. This allowed us to attribute systematic differences in gene expression between the two groups to environmental influences. To test the effect of population of origin we grouped the samples halves based on the site they originated from regardless of their transplant location during the study period. This allowed us to test for differences in gene expression that were distinctive of the two populations irrespective of environment. Expression data for each sample and estimates of differential expression between sample groups were uploaded through Zenodo at the following address [DOI](https://zenodo.org/badge/doi/10.5281/zenodo.12626.png)](http://dx.doi.org/10.5281/zenodo.12626) and will be available after June 1st 2015 or earlier. The read files have been uploaded the NCBI SRA (accession SRP049522) and will be released by November 11th 2015 or earlier.

When sample halves were grouped based on transplantation site, 321 genes showed significant differential expression between the two groups (adjusted $p < 0.01$). We refer to this set of differentially expressed genes as 'environmentally flexible genes', since they are regulated depending on which site the sample halves were placed at. Environmentally flexible genes were

significantly over-represented within the high-CpG component (Figure 7A and Table 1). Thus genes showing signatures of weak germline methylation were more likely to display environmentally driven variation in expression. To assess this relationship on a continuous scale we plotted $CpG_{o/e}$ against the magnitude of differential expression for each gene, calculated as [log(mean expression in environment A / mean expression in environment B)]. Differential expression was positively correlated with $CpG_{o/e}$ (Spearman's rank correlation; rho = 0.138; $p << 0.0001$; Figure 7B). To examine if this was a simple linear relationship genes were divided into twelve quantiles based on $CpG_{o/e}$ and mean differential expression was plotted for each quantile. This analysis revealed that the magnitude of differential expression increased sharply within the high-$CpG_{o/e}$ component (Figure 7C), suggesting that differential expression correlates categorically with $CpG_{o/e}$ component rather than continuously with increasing $CpG_{o/e}$.

Link between CpGo/e and population-specific gene expression

As differential expression due to transplantation site showed strong bias towards high $CpG_{o/e}$, we performed the same analyses on differential expression with respect to sample origin. Here we compared gene expression between sample halves grouped based on their site of origin. We found 68 genes that maintained origin-specific expression patterns (adjusted $p < 0.01$) irrespective of transplantation site. Differential expression due to sample origin showed similar positive associations with $CpG_{o/e}$ (Figure 7D-F).

CpGo/e and gene expression level

To investigate broad scale relationships between the magnitude of gene expression and $CpG_{o/e}$, we plotted mean transcript abundance (across all samples) for each gene against its $CpG_{o/e}$ value (Figure 8). Genes were divided into 25 equally sized quantiles based on $CpG_{o/e}$. Mean transcript abundance varied significantly across $CpG_{o/e}$ quantiles (ANOVA $p << 0.0001$) and genes

in the high-$CpG_{o/e}$ component showed decreased mean expression compared genes in the low-$CpG_{o/e}$ component (Welch Two Sample t-test; $p << 0.0001$).

DISCUSSION

Bimodal patterns of gene body methylation as an ancestral feature among Metazoa

Depletion of CpG dinucleotides, a signature for historic germline DNA methylation, is widespread in *A. millepora* and follows a characteristic bimodal pattern. Notably, even for the high $CpG_{o/e}$ component, the mean $CpG_{o/e}$ value of 0.74 was less than 1.0, suggesting that these genes also bear signatures of germline methylation, although apparently weaker than genes in the low-$CpG_{o/e}$ component. As shown by Sarda *et al*. (Sarda et al. 2012), bimodal methylation is consistent among diverse invertebrate taxa: reported in Hymenoptera (Park et al. 2011)(Elango et al. 2009), Hemiptera (Hunt and Brisson 2010), Lepidoptera (Xiang et al. 2010), Orthoptera (Falckenhayn et al. 2013), Mollusca (Gavery and Roberts 2010), and Cnidaria (Zemach et al. 2010; Sarda et al. 2012; This study). Evidence of bimodal methylation in Cnidaria (the sister group to all bilaterians) along with other diverse taxa suggests an ancient mechanism that has been conserved through more than 500 million of years of evolution (Chapman et al. 2010).

Correlation between CpGo/e and Gene Function in *A. millepora*

Consistent with previous findings among diverse invertebrates (Sarda et al. 2012; Elango et al. 2009), we found significant variation in $CpG_{o/e}$ between different biological processes. The distribution follows a characteristic trend in which functions with spatial and temporal stability are enriched in the low-$CpG_{o/e}$ component, implying strong methylation. Functions with greater spatial and temporal variability are enriched in the high-$CpG_{o/e}$ component, implying weak methylation. Thus, our results add to previous findings and suggest an association between weak gene body methylation and dynamic biological processes. In light of these results it was surprising that the GO term *stress response* showed an intermediate mean $CpG_{o/e}$ and only slight enrichment in the

high-$CpG_{o/e}$ component (Figure 6). This contrasts with results from Gavery and Roberts (Gavery and Roberts 2010), where stress response genes showed elevated $CpG_{o/e}$. It is possible that the intermediate value for *stress response* in our system is related to regularity with which reef-building corals must contend with cellular stress. Unlike to other invertebrates examined for $CpG_{o/e}$ previously, scleractinian corals are host to photosynthetic endosymbionts (*Symbiodinium* spp.) (Rocker et al. 2012). While the coral host depends on endosymbionts for fixed carbon, their daily photosynthetic activity causes a hyperoxic cellular environment, so that host cells are regularly exposed to oxidative stress (Levy et al. 2011). The relatively low mean $CpG_{o/e}$ for stress response in *A. millepora* could possibly reflect historical methylation of cellular stress response genes that require highly regular expression patterns to mediate chronic stresses of symbiosis. To support this, we compared mean $CpG_{o/e}$ of GO terms nested within stress response and found that *oxidative stress*, *response to wounding*, and *cellular response to stress* had the lowest mean $CpG_{o/e}$ values (Figure 9). A recent comparison of gene expression in sea anemone (*Aiptasia*) revealed that along with *response to oxidative* stress, genes involved in *Inflammation, tissue remodeling*, and *response to wounding*, showed strong differential expression between symbiotic and asymbiotic individuals (Lehnert et al. 2014). After removal of genes associated with *response to oxidative stress* and *cellular response to stress*, the remaining 426 stress response genes were substantially enriched in the high-$CpG_{o/e}$ component (Fisher's exact test $p < 0.001$). These results suggest the interesting possibility that invertebrates possess taxon-specific methylation patterns based on the demands of their particular life histories.

Link between CpGo/e and gene expression plasticity

In insects, weakly methylated status is linked with variation in gene expression across tissue types (Foret et al. 2009), caste (Elango et al. 2009), and developmental stages (Wang et al. 2013), and a similar trend across tissues types been shown in oysters (*Crassostrea gigas*) (Gavery and Roberts 2013). However, whether the same association is found regarding variation in gene expression across different habitats remained unresolved. Here we show that in *A. millepora*, genes

with high $CpG_{o/e}$ values are more likely to be differentially expressed in response to environmental conditions (Figure 7A). High-$CpG_{o/e}$ genes also show greater magnitude of expression change in response to the environment (Figure 7B-C). To the extent that $CpG_{o/e}$ scores correlate with gene body methylation in somatic cells at the time of sampling, these results suggest a link between weak gene body methylation and gene expression plasticity.

Roberts and Gavery (Roberts and Gavery 2012) proposed a role for gene body methylation in modulating plasticity in which weak gene body methylation passively facilitates flexible gene expression via alternative transcription start sites (TSSs), increased sequence mutations, exon skipping, or through transient gene methylation. Our results are consistent with the general hypothesis that weak gene body methylation facilitates environmentally responsive expression. To offer a potential mechanism, weak methylation could increase expression plasticity by allowing greater access to alternative TSSs. Among developmental genes in *Drosophila melanogaster*, alternative promoter use is common and can mediate temporally regulated expression (Batut et al. 2013). In mammals, alternative promoter use is associated with tissue specific expression and can be regulated though methylation of intragenic CGIs (Maunakea et al. 2010). Alternative TSSs could respond differently to regulatory elements and environmental cues, thereby permitting more complex and responsive transcriptional regulation. As proposed by Elango et al. (Elango et al. 2009), high $CpG_{o/e}$ genes also could be more amenable to epigenetic modulation. Here genes that are generally weakly methylated could be conditionally methylated to mediate environmentally responsive expression. This mechanism would be of particular interest if methylation patterns were shown to fluctuate in response to environmental cues. Because $CpG_{o/e}$ is an evolutionary signature reflecting historical methylation rather than the particular state at the time of sampling, this is a possibility that our approach could not address. An alternative hypothesis is that gene body methylation has little or no regulatory effects and is simply a byproduct of transcriptional patterns mediated by other mechanisms. Further experimentation will be required to resolve the potential explanations.

Link between CpGo/e and population-specific expression

We also identified genes with differential expression based on the corals' origins. These genes showed expression patterns that were distinctive of the two populations and robust to changes in environmental conditions (i.e. expression differences between populations were maintained irrespective of transplant site). Our initial expectation was to find enrichment of low-$CpG_{o/e}$ genes among these stably expressed origin-specific genes; however, we found that these genes tended toward high $CpG_{o/e}$ similarly to the environmentally flexible genes (Figure 7D-E).

Since previous studies have demonstrated genetic structure between the Orpheus and Keppel populations (Van Oppen et al. 2011), the simplest explanation of the origin-specific expression differences is that they result from genetic divergence. If this is the case, an intriguing possibility is that gene body methylation may help to stabilize expression patterns across divergent genetic backgrounds. Such buffering against genetic variation would be similar to the function of the heat shock protein 90 (Rohner et al. 2013), except at the level of transcription rather than at the level of protein structure.

CpGo/e shows negative relationship with mean expression level

Although the difference was subtle, genes in the low-$CpG_{o/e}$ component (indicating strong germline methylation) showed higher average expression (Figure 8). This could be attributed to the ubiquitous expression of low-$CpG_{o/e}$ "housekeeping" genes, whereas the expression of many high-$CpG_{o/e}$ genes is restricted to certain cell types and thus might appear lower on the scale of the whole organism. It is also possible that gene body methylation lowers gene expression noise, as has been shown in human blood and brain tissue (Huh et al. 2013), potentially leading to less frequent "off" state and thus higher average expression of low-$CpG_{o/e}$ genes. A similar trend has been shown in the yeast (*Saccharomyces cerevisiae*) where transcriptional noise is negatively associated with protein abundance, and genes involved responding to the environment show elevated noise compared to genes involved in protein synthesis (Newman et al. 2006).

The link between predicted methylation and elevated expression contrasts with results from Riviere et al. (Riviere et al. 2013), in which methylation of homeobox genes of oysters during development was inversely related to mRNA abundance. They suggest that in the case of these developmental genes a 'CpG-island-like' mechanism of gene repression could be responsible decreasing mRNA abundance. Hence it does not seem that gene body methylation leads to increased expression *per se* and the trend we describe is more likely due to enrichment with ubiquitously expressed housekeeping genes and not to a methylation-driven increase in expression of low-$CpG_{o/e}$ genes.

Corals as a model to study ecological roles of gene body methylation

Reef-building corals, as phylogenetically basal metazoans with a well-studied ecology and emerging genomic resources, represent an excellent study system to address the function of gene body methylation. Unlike most other animal models, corals can be fragmented into clonal replicates and transplanted across natural environments, making it possible to disentangle environmental and genotypic effects on genome-wide processes in realistic ecological contexts (Kenkel et al. 2013). Corals can also be crossed in the lab to generate full-sib families of larvae and juveniles, which enables studies of how gene body methylation becomes established and facilitates quantitative genetic analysis (Meyer et al. 2009). Lastly, understanding the role of DNA methylation in phenotypic plasticity of corals has important conservation implications. Phenotypic plasticity is predicted to significantly influence evolutionary responses to climate change (Reusch 2013) and corals, the foundation of the most biologically diverse ecosystem in the ocean, are among the species most vulnerable to extinction (Foden et al. 2013).

CONSLUCIONS

Our results indicate a connection between historical germline methylation and gene expression flexibility across environments and populations. As a whole our results are consistent

with a hypothesis that strong gene body methylation leads to more stable gene expression while weak methylation facilitates flexible expression, although the direction of causality remains to be confirmed. Studies of phylogenetically basal invertebrates such as corals will further elucidate the fundamental functional aspects of gene body methylation in Metazoa.

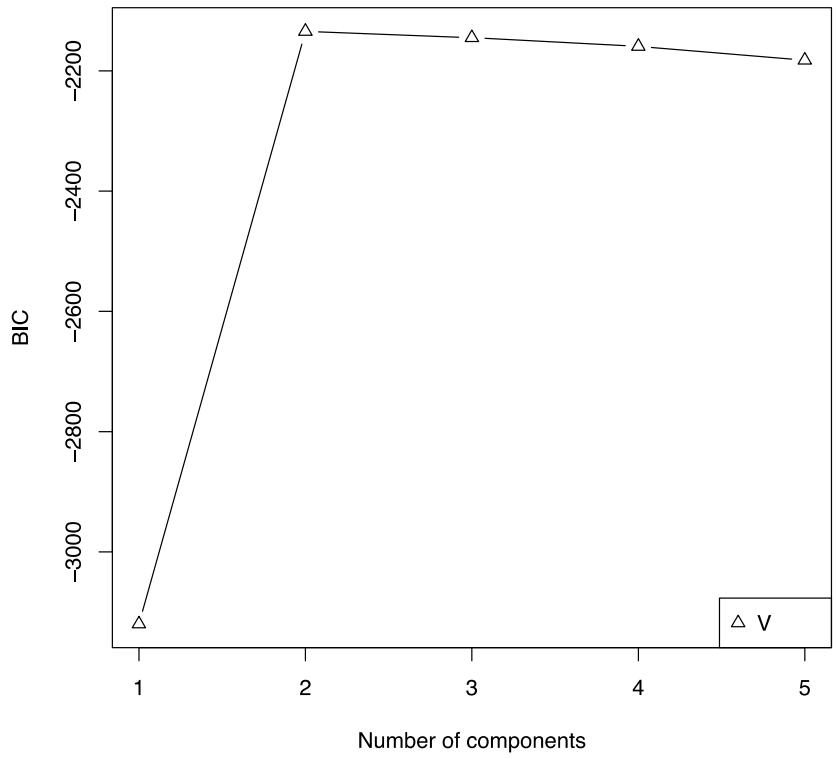

Figure 4 Estimated fit for Gaussian Mixture Models: Bayesian information criterion (BIC) was used to compare the fit of Gaussian Mixture Models with different numbers of components to the distribution of CpGo/e values. BIC indicated that a two-component model provided better fit than a single component model.

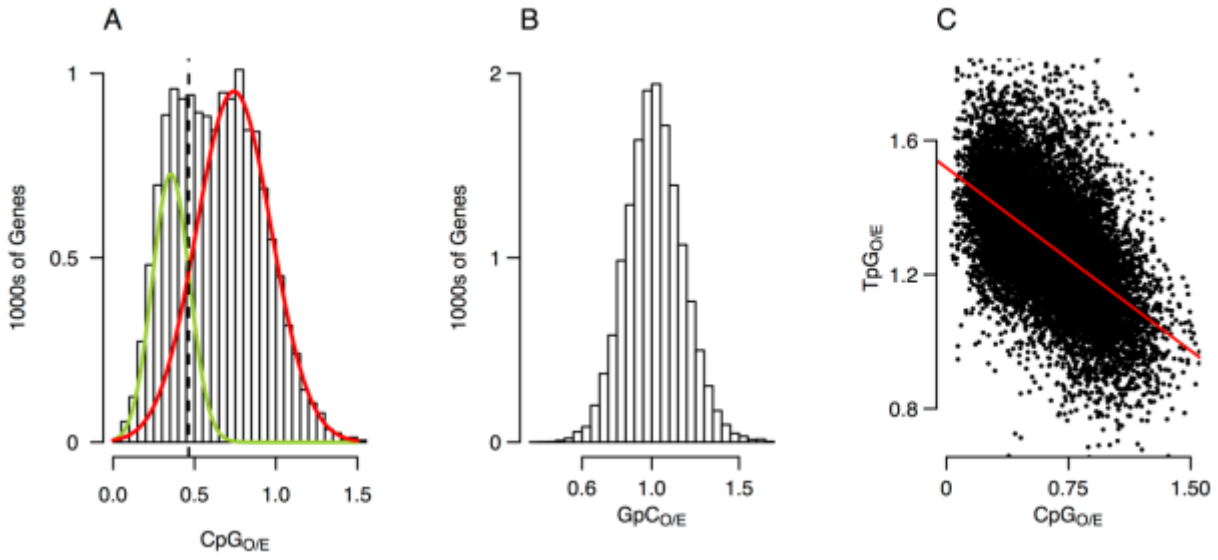

Figure 5 Signatures of gene body methylation are bimodally distributed in the coral. (A) Distribution of genes based on normalized CpG content. The green curve indicates the low-CpG component (predicted to be strongly methylated). The red curve indicates the high-CpG component (predicted to be weakly methylated). The black dotted line separates the two components at the point of intersection between the curves. (B) Distribution of genes based normalized GpC content. In contrast to CpG dinucleotides, GpCs are not targeted for methylation so a normal distribution is expected. (C) Negative linear relationship between $CpG_{o/e}$ and $TpG_{o/e}$. This is consistent with the prediction that DNA methylation causes depletion of CpG content largely though substitution of methylated cytosines for thymine.

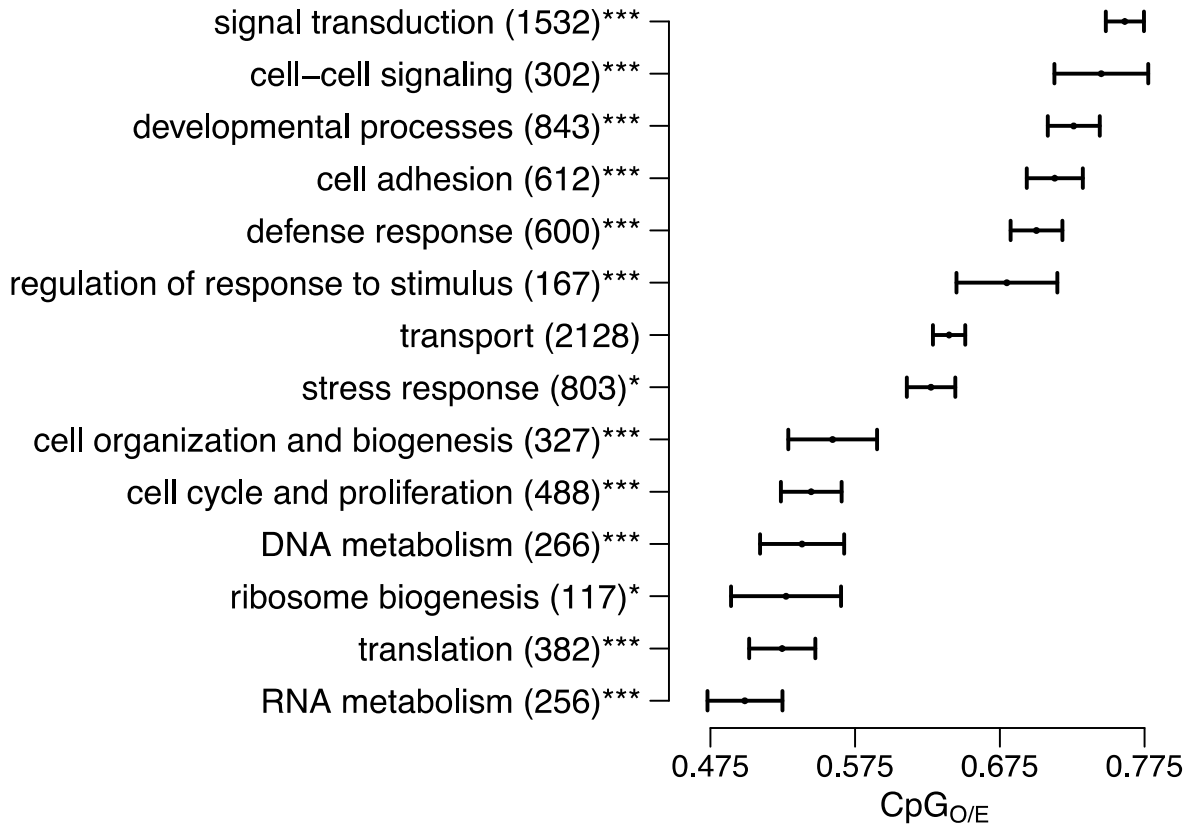

Figure 6 Variation of $CpG_{o/e}$ among genes assigned to different biological processes. Each bar represents mean $CpG_{o/e}$ for the indicated biological process and its standard error. Asterisks indicate significance of enrichment in the low- or high-CpG components (*< 0.05, **< 0.01, ***< 0.001; Fisher's exact test).

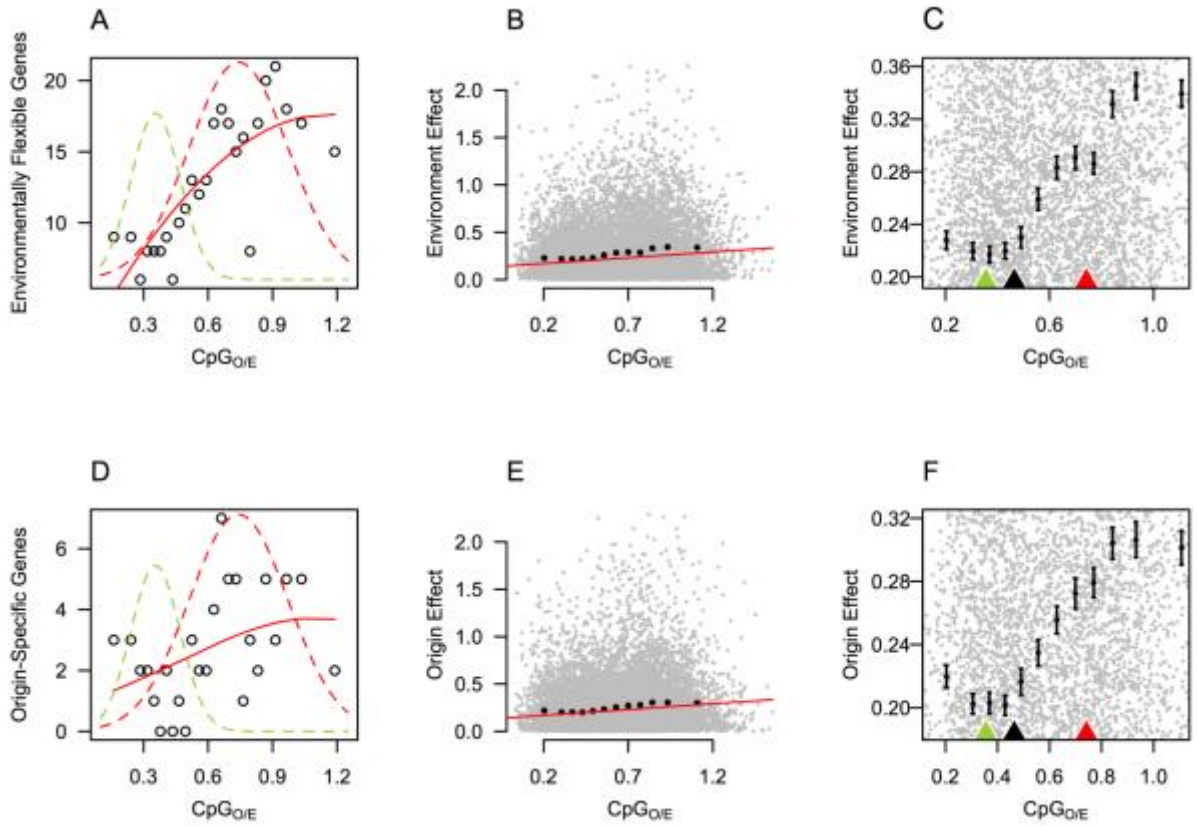

Figure 7 Genes with high $CpG_{o/e}$ are more likely to be differentially expressed between environments. (A) Frequency of environmentally flexible genes increases with $CpG_{o/e}$. All genes with expression data were divided into 25 quantiles based on $CpG_{o/e}$ (503 genes per quantile). Each data point represents the count of environmentally flexible genes (adjusted *P*-value < 0.01) within a single quantile and the mean $CpG_{o/e}$ for the quantile. To illustrate associations with the $CpG_{o/e}$ components, the density component curves from Figure 5A were traced over the count data. (B) Across all genes the magnitude of differential expression due to environment (environment effect) showed a positive relationship with $CpG_{o/e}$. The red line indicates the linear model of the relationship between environmental effect and $CpG_{o/e}$. Black error bars represent the mean and standard error for environmental effect of 12 quantiles based on $CpG_{o/e}$. (C) Same as (B), rescaled to illustrate that mean environment effect increases sharply under the high-CpG component. Green and red arrows along the x-axis illustrate the means for each component curve. The black arrow indicates their point of intersection. (D-F) Same as A-C, but for the effect of coral origin rather than of transplant site.

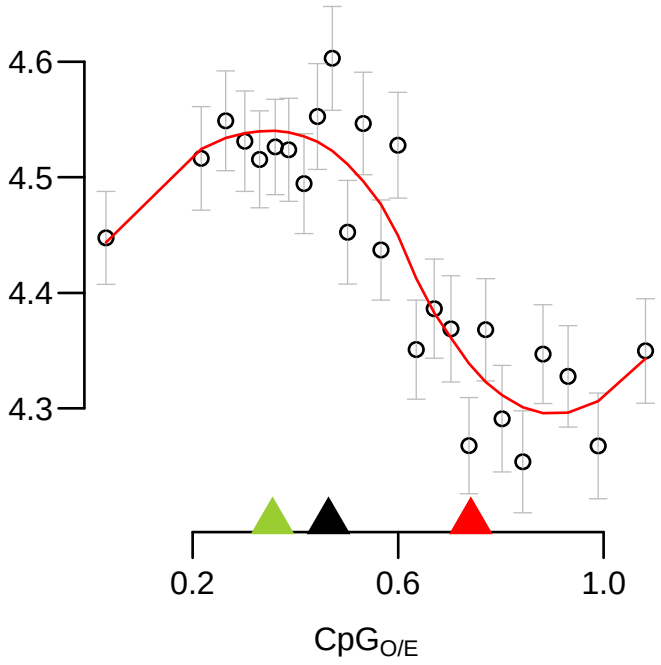

Figure 8 Correlation of $CpG_{o/e}$ with transcript abundance. Mean gene expression values were generated from 25 equally sized quantiles based on $CpG_{o/e}$. Each gene was assigned an expression value equal to its average expression across all samples. Each data point represents mean of the expression values for all genes included in the quantile plotted against mean $CpG_{o/e}$ for the quantile; the whiskers denote standard errors. Green and red arrows indicate the means for the two mixture component shown in Figure 5A. The black arrow indicates the point of separation between the components.

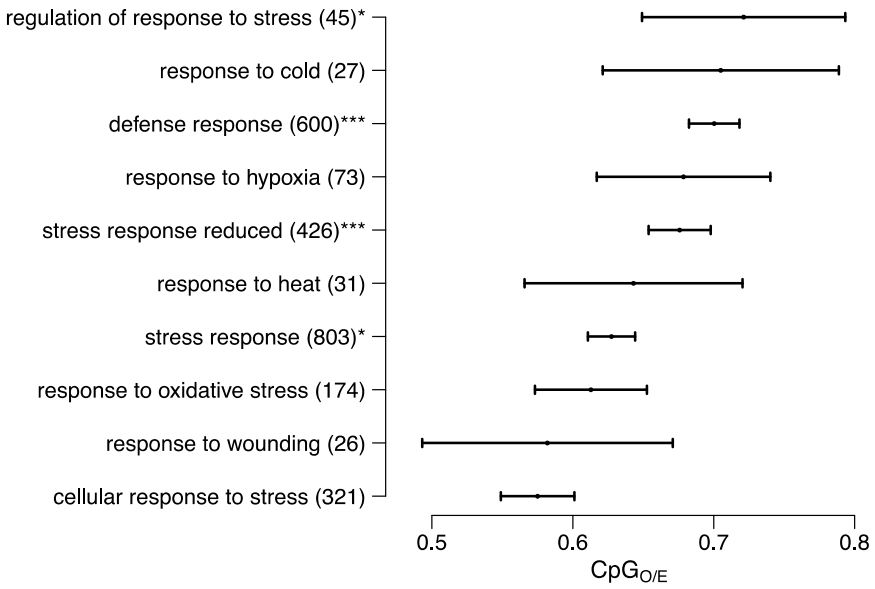

Figure 9 Genes involved in *response to oxidative stress* and *cellular response to stress* contribute to the relatively low mean $CpG_{O/E}$ for the *stress response* Gene Ontology term. The figure illustrates the variation in $CpG_{O/E}$ of Gene Ontology (GO) terms nested within stress response. Each bar represents mean $CpG_{O/E}$ for the indicated GO term and its standard error. Asterisks indicate significance of enrichment in the low- or high-CpG components (* < 0.05, ** < 0.01, *** < 0.001; Fisher's test). The bar labeled 'stress response reduced' represents the *stress response* GO term with genes from *response to oxidative stress* and *cellular response to stress* removed. GO terms with fewer than 20 representative genes were not plotted.

Chapter 2b: Evolutionary consequences of gene body methylation in *Acropora millepora*

ABSTRACT

Gene body methylation (GBM) is an ancestral and widespread feature in Eukarya, yet its adaptive value and evolutionary implications remain unresolved. The occurrence of GBM within protein coding sequences is particularly puzzling, because methylation causes cytosine hypermutability and hence is likely to produce deleterious amino acid substitutions. We investigate this enigma using an evolutionarily basal group of Metazoa, the stony corals (order Scleractinia, class Anthozoa, phylum Cnidaria). We show that patterns of coral GBM are similar to other invertebrate species, predicting wide and active transcription and slower sequence evolution. We also find a strong correlation between GBM and codon bias, resulting from systematic replacement of CpG bearing codons. We conclude that GBM has strong effects on codon evolution and speculate that this may influence establishment of optimal codons.

INTRODUCTION

DNA methylation is an evolutionarily widespread epigenetic modification found in plants, animals and fungi. It is defined as the covalent addition of a methyl group to the one of the four DNA bases, predominantly on the fifth carbon of cytosines within CG dinucleotides (CpGs), producing 5-methylcytosine (5mC). Unlike other epigenetic modifications, DNA methylation not only alters chromatin structure and transcription, it changes the mutation rate of the underlying DNA. This is because 5mC undergoes deamination reactions more readily than normal cytosine (Shen et al. 1994) and deamination produces thymine rather than uracil, which is less likely to be accurately repaired (Zemach and Zilberman 2010). Because of this hypermutability, sequences that are heavily methylated in the germ-line become deficient in CpGs, with corresponding increases in TpGs and CpAs (Sved and Bird 1990). Hence DNA methylation has evolutionary consequences outside of its direct physiological effects.

Evolutionary effects of 5mC hypermutability are apparent in both vertebrate and invertebrate genomes. In mammals, DNA methylation is ubiquitous, so that nearly all genomic regions show lower than expected frequency of CpGs (Karlin and Mrázek 1996; McGaughey et al. 2014). The exception is regions of elevated CpG content called CG islands that are protected from DNA methylation (Jones 2012). In most invertebrates, DNA methylation is not ubiquitous but patchy, occurring primarily on CpGs within gene bodies (Suzuki et al. 2007; Zemach et al. 2010). This intragenic form of DNA methylation is referred to as gene body methylation (GBM). In invertebrate genomes, GBM occurs preferentially on actively and widely expressed genes, resulting in covariations between genes' CpG content, function, and expression patterns (Elango et al. 2009; Hunt and Brisson, 2010; Zemach et al. 2010; Sarda et al. 2012). Similar patterns of genic methylation are found in plants (Tran et al. 2005; Zilberman et al. 2007) and mammals (Baubec et al. 2015).

Despite this widespread phylogenetic occurrence, GBM is by no means universal. In several groups, such as yeast (*Saccharomyces cerevisiae)*, fruit fly (*Drosophila melanogaster*), worm (*Caenorhabditis elegans*), and the basal plant *Marchantia polymorpha*, genic methylation is extremely scarce or lost altogether (Capuano et al. 2014; Takuno et al. 2016). It has been proposed that the secondary loss of DNA methylation occurs because its mutational costs outweighed its adaptive value (Zemach *et al.* 2010). Indeed, even within gene bodies, methylation occurs preferentially on exons (Zemach *et al.* 2010; Libbrecht et al. 2016), where mutations are likely to have the greatest deleterious effect. In humans, genic methylation increases deleterious *de novo* mutations with paternal age (Francioli et al. 2015). Why, given its apparently nonessential and outright mutagenic nature, has GBM persisted for so long across such a diversity of taxa?

In this study, we investigate the evolutionary consequences of 5mC on invertebrate coding sequences. Using the first direct genome-wide characterization of DNA methylation in a reef-building coral, we confirm previous studies showing that GBM predicts active and stable gene expressive (Elango et al. 2009; Hunt and Brisson, 2010; Zemach et al. 2010, Sarda et al. 2012). We also test previous findings that in spite of 5mC hypermutability, GBM predicts slower

sequence evolution (Park et al. 2011; Takuno and Gaut 2012). Finally, we examine GBM in the context of synonymous codon usage. Because GBM occurs preferentially on a subset of invertebrate coding genes (Sarda et al. 2012), we hypothesized that its mutagenic effects cause intragenomic variation in codon bias. While methylation is often cited as an explanation for patterns of codon usage (Kanaya et al. 2001; Sterky et al. 2004; Gonzalez-Ibeas et al. 2007; Qin et al. 2013; Duan et al. 2015), direct investigations of this hypothesis have been lacking. As a basal metazoan predicted to have a typical bimodal pattern of GBM (Dixon et al. 2014; Dimond and Roberts 2016), the branching coral *Acropora millepora* was well suited to address this problem.

METHODS

Sequence Data and Computational Tools

Transcriptomic data from 17 species of Scleractinia (stony corals) and 3 species of Actiniaria (anemones) were downloaded from the web (Table 1; Schwarz et al. 2008; Sunagawa et al. 2009; Polato et al. 2011; Shinzato et al. 2011; Moya et al. 2012; Kenkel et al. 2013; Lubinski and Granger 2013; Sun et al. 2013; Maor-Landaw et al. 2014; Nordberg et al. 2014; Willette et al. 2014; Kitchen et al. 2015; Davies et al. 2016)). Instructions, scripts, and example output files for computational methods used in this study are available on GitHub (https://github.com/grovesdixon/metaTranscriptomes). Gene Ontology and KOG annotations were applied as described in (Dixon et al. 2015). Instructions and scripts for the gene annotation pipeline are available on GitHub (https://github.com/z0on/annotatingTranscriptomes). Significance for enrichment of KOG terms across MBD-scores was tested using Mann-Whitney U tests implemented in the R package KOGMWU as in Dixon et al. (2015).

Ortholog Identification and alignment

Orthologs were identified based on reciprocal best Blast hits between extracted protein sequences. First, coding and amino acid sequences were extracted from each transcriptome based

on alignments (e-value cutoff = 1e-5) to a reference proteome using BlastX (Altschul et al. 1997) and a custom Perl script CDS_extractor_v2.pl (https://github.com/z0on/annotatingTranscriptomes), which identifies and corrects frame shift mutations within the BlastX-aligned sequences. The reference proteome was a concatenation of the *Nematostella vectensis* (Nordberg et al. 2014) and *Acropora digitifera* (Shinzato et al. 2011) reference proteomes. The protein sequences for all pairs of species were reciprocally blasted using BlastP (Altschul et al. 1990). Because our MBD-seq dataset was generated from *A. millepora*, we used its sequences as anchors for orthologous groups. First an initial set of candidate orthologs was compiled based on reciprocal best hits between *A. millepora* and each other species. Only hits with alignment lengths >75% of the subject sequence and an e-value < 1e-5 were retained. This initial set was then refined to include only sequences that were reciprocal best hits with >= 50% of other candidate orthologs within the group (Figure 10). Orthologous groups with fewer than three (15%) representative species were excluded. For building the species tree, a separate, highly conserved set of orthologs was assembled with percent amino acid identity > 75%. These were further filtered by retaining only orthologs with representative sequences from > 80% of species. As a final filter, we used cluster analysis of dS values to identify likely paralogs and spurious orthologs. For each species, a three component Gaussian mixture model was fit to the pairwise dS estimates with *A. millepora*. The first two components were assumed to capture the true orthologs, the third component was assumed to have captured false positive orthologs (Figure 11). Mean dS for the third component was on average 60 times higher than the second highest component. On average 10% of ortholog calls were flagged as false positives and removed. Amino acid sequences for each ortholog were aligned with MAFFT (Katoh and Standley 2013) using the 'localpair' algorithm. The protein alignments were then reverse translated into codon sequences using Pal2Nal (Suyama et al. 2006).

Substitution rate analyses

To estimate substitution rates (dS and dN) we used codeml in the software package PAML (Yang 2007). Substitution rates were estimated using pair-wise comparisons between *A. millepora* and each other species that had representative sequences for each ortholog. Example codeml control files for the pair-wise comparisons are available on GitHub (https://github.com/grovesdixon/metaTranscriptomes).

Building species tree

Based on a highly conserved set of ortholog sequences we constructed species tree using RAxML (Stamatakis 2014). For phylogenetic construction, we ran the rapid bootstrapping algorithm using the GTRGAMMA model and 1000 iterations. We decided to use putative orthologs with representative sequences in > 80% of taxa through iterations of tree building. The best trees from ortholog sets using 40%, 50% and 60% cutoffs all gave the same topology. The best tree using the 80% cutoff was chosen for because it had highest bipartition bootstrap values.

Library preparation for MBD-seq

To quantify GBM in *Acropora millepora* we used methyl-CpG binding domain protein-enriched sequencing (MBD-seq). Enrichment reactions were performed using the MethylCap kit (Diagenode Cat. No. C02020010). Seven enrichment reactions were performed. Input DNA for all reactions was isolated from a single colony of *A. millepora* sampled from the Central Great Barrier Reef (Great Barrier Reef Marine Park Permit G09/29894.1). DNA was diluted to 0.1 µg/µl then sheared with a Misonix Sonicator 3000 for nine or ten minutes using 15 second cycles at ~30W. Sheared fragments ranged from ~100 to 800 bp spanning the range 300 – 500 bp recommended by the manufacturer. The manufacturer's protocol recommended an input of 12 µl of sheared DNA diluted in 130 µl of buffer. We found that our yields were higher when we used 18, 24 or 48 µl of sheared DNA (1.5x, 2x and 4x concentrated). As the kit is intended for mammalian DNA, lower genome wide methylation levels in our system could explain why higher input concentrations

worked better in our case. Flow-through from initial capture of methylated DNA was retained for sequencing. The methylated fraction was eluted from the capture beads in a single step using High Elution Buffer. Electrophoresis gels were used to assess the size and quality of each elution. For sequencing, product from enrichment replicates 1-4 and 5-7 were pooled. Final concentrations of these pooled libraries were 6 and 4 ng/µl measured with a Nanodrop Spectrophotometer. Similarly, the flow-through components from the same replicates were pooled. Concentrations of the flow-through pools were 34 and 36 ng/µl. Adapter ligation using a NEBnext kit (New England Biolabs®), library quality assessment using a Bioanalyzer (Agilent Technologies), and sequencing on a HiSeq 2500 platform (Illumina®) were performed by the University of Texas Genome Sequencing and Analysis Facility.

Analysis of gene body methylation

Raw reads from the MBD-sequencing libraries were trimmed using cutadapt (Martin 2011) and quality filtered using Fastx toolkit (http://cancan.cshl.edu/labmembers/gordon/fastx_toolkit/). Reads were then aligned to coding sequences extracted from the *A. millepora* reference transcriptome (Moya et al. 2012), as described above. DESeq2 (Love et al. 2014) was used to calculate the $\log_2$ fold difference between the MBD-enriched and flow-through libraries. We used this $\log_2$ fold difference, which we refer to as MBD-score, as our quantification of the strength of GBM for each gene. Negative values indicate weak methylation and positive values indicate strong methylation. To examine the distribution of MBD-scores we used the R package Mclust (Fraley and Raftery 2007). We first assessed the optimal mixture model and number of components based on Bayesian Information Criterion (BIC). The optimal number of components was greater than one with little change in BIC beyond two components (Figure 12A). Based on this result we fitted a two-component mixture model to the MBD-scores (Figure 12B).

Because of the hypermutability of 5mC, genes that are strongly methylated in the germline become deficient in CpG dinucleotides over evolutionary time (Sved and Bird 1990). As a result, normalized CpG content (CpGo/e) can be used to estimate historical germline methylation. This

metric has been shown to correlate closely with direct measures of GBM (Zemach et al. 2010; Sarda et al. 2012). To corroborate that our measure of GBM also correlated with CpGo/e we calculated it for the *A. millepora* coding regions as described in Dixon et al. (2014). To control for effects on gene length, CpGo/e was calculated based on the first 1000 bases of each sequence.

Gene expression datasets

To test for correlations between MBD-score and transcriptional variation we used gene expression data from two previous experiments. Both datasets were generated using Tag-based RNA-seq (Meyer et al. 2011) from samples of *A. millepora* taken from the Central Great Barrier Reef, Australia. The current laboratory and bioinformatics protocols for analysis of Tag-based RNA-seq are available on GitHub (https://github.com/z0on/tag-based_RNAseq). The first dataset was a subset of that described in (Dixon et al. 2015), including 12 adult samples (triplicate samples from 2 genotypes from Princess Charlotte Bay and 2 from Orpheus Island: Great Barrier Reef Marine Park Authority permit G38062.1 exposed to 28°C) and 30 samples of their larval offspring (10 genetic families, reared for five days at 28°C in triplicate). Variation in gene expression between adults and larvae was analyzed using DESeq2 (Love et al. 2014). Comparisons between MBD-score and transcript abundance were based on counts from adult samples transformed to a $\log_2$ scale using the rlog Transformation function. Mean expression levels from this dataset were also used to calculate indices of codon bias described below. The second dataset described in (Dixon et al. 2014) included 56 colony fragments reciprocally transplanted between two environmentally distinct reefs: Keppel and Orpheus Island (Keppel: 23°09S 150°54E and Orpheus 18°37S 146°29E: Great Barrier Reef Marine Park Authority permit G09/29894.1). Expression profiles from these samples were analyzed with respect to the transplantation site to examine variation in gene expression due to environmental conditions.

Analysis of codon bias

We tested for relationships between MBD-score and synonymous codon usage using four metrics: relative synonymous codon usage (RSCU), frequency of optimal codons (Fop), codon adaptation index (CAI), and the effective number of codons (Nc). RSCU was calculated as the ratio of the observed number of occurrences of a particular codon to the excepted number of occurrences if codon usage was neutral (Sharp et al. 1986):

$$RSCU_{ij} = \frac{x_{ij}}{1/n_i \sum_{j=1}^{n_i} x_{ij}}$$

Where X_{ij} is the number of occurrences of the *j*th codon for the *i*th amino acid and n_i is the number of synonymous codons for the *i*th amino acid. This measure quantifies relative codon usage while controlling for variation in amino acid composition between proteins. Fop is intended to measure the degree of selection for optimal codon usage in a particular coding sequence. It was originally defined as the ratio of optimal codons to the total number of codons in a gene, with optimal codons identified based on the cellular content of isoaccepting tRNAs and the nature of codon-anticodon interactions (Ikemura 1981). Optimal codons are also inferred based on relative usage in a set of highly expressed genes such as ribosomal proteins (Behura and Severson 2013). To estimate Fop for *A. millepora* coding regions we used the software package CodonW (Peden 1999)(http://codonw.sourceforge.net/). CodonW uses correspondence analysis of codon usage to derive a set of optimal codons and then estimates their usage for each sequence. CAI is similar to Fop, and is intended to quantify the strength of selection on codon usage. For a given gene, CAI is equal to the geometric mean of the relative adaptiveness (*W*) of all codons within that gene. The relative adaptiveness W_{ij} of codon *i* that codes for amino acid *j* is equal to the ratio its relative synonymous codon usage to that of the most abundant synonymous codon in a set of highly expressed genes (Sharp and Li 1987a):

$$W_{ij} = RSCU_{ij}/RSCU_{imax}$$

Relative adaptiveness (based on the top 5% most highly expressed genes) and CAI were calculated using custom python scripts. Unlike CAI and Fop, Nc does not depend on a set of preferred codons,

and provides an estimate of a gene's departure from random use of synonymous codons based solely on codon usage. The measure is analogous to the 'effective number of alleles' in population genetics (treating amino acids as loci and codons as alleles). For a given coding sequence, Nc is the effective number of codons summed across the 20 amino acids. It and is bounded between 20 (completely biased) to 61 (neutral)(Wright 1990). Nc was calculated using CodonW (Peden 1999)(http://codonw.sourceforge.net//culong.html).

Statistical Analyses

Statistical analyses of the relationship between MBD-score and other gene characteristics were performed in R (R Core Team 2015). Significance for correlations was established using Spearman's rank-order correlation test. Significance tests for differences in counts between the strongly methylated and weakly methylated classes were performed using Fisher's exact tests (Fisher 1922). Principal component analysis was performed using prcomp function in R.

RESULTS

Using MBD-seq to quantify gene body methylation

We used Methylation Binding Domain enrichment sequencing (MBD-seq)(Harris et al. 2010) to measure GBM in *A. millepora*. The strength of methylation for 24320 coding regions was quantified as the $\log_2$ fold difference between captured and flow-through fractions of MBD enrichment preparations. We refer to this $\log_2$ fold difference as the MBD-score. Raw read data are publicly available through the NCBI SRA database (SRA accession: SRP074615). Analysis of the distribution of MBD-scores (Figure 13A) showed that it was best described as a mixture of two or more Gaussian components (Figure 12). MBD-score correlated with CpGo/e, indicating that our measure of GBM overlapped closely with historical patterns of germ-line methylation (Figure 13B). As an MBD-score of zero indicated equal representation in the captured and flow-through fractions we chose this value to separate strongly and weakly methylated genes. Genes with MBD-

scores greater than zero are referred to as strongly methylated genes, those with scores less than zero are referred to as weakly methylated.

MBD-score is linked with gene function and expression patterns

MBD-score was associated with gene function. Analysis of selected GO categories for biological processes revealed that strongly methylated genes tend toward biological functions that are spatially and temporally stable, such as DNA metabolism, ribosome biogenesis, translation, RNA metabolism and transcription. Weakly methylated genes tended to involve biological functions that are spatially and temporally regulated, such as cell-cell signaling, response to stimulus, signal transduction, cell adhesion, defense response and development (Figure 14A). Clustering of KOG categories for higher or lower MBD-scores further supported these results (Figure 14B).

To test if weak GBM is a signature for inducible transcription we correlated MBD-score with RNA-seq data, comparing different developmental stages and environmental conditions. For developmental stage, $\log_2$ fold differences in transcript abundance between *A. millepora* adults and larvae (described in Dixon *et al.* 2015) were negatively correlated with MBD-score (Figure 15A). Significantly differentially expressed genes (DEGs at FDR < 0.01) were 1.4 times more frequent among weakly methylated genes (Figure 15B). A similar trend was found for variation in expression due to environmental conditions (Figure 15C-D). Here clonal fragments of adult colonies were exposed to two environmentally distinct regimes for three months prior to sampling for RNA-seq (Dixon *et al.* 2014). Differential expression (FDR < 0.01) between environmental regimes was 2.2 times more frequent among weakly methylated genes.

MBD-score also showed weak but significant correlation with transcript abundance (Figure 16A and B). Highly expressed genes were on average strongly methylated (Figure 163C and D). The top 5% most strongly methylated genes however, showed lower average expression (Figure 16E). This indicates that while GBM is generally associated with elevated transcription, extreme

levels may be inhibitory. This appears to be particularly true for short genes, as the removal of coding sequences shorter than 800 bp mitigated the trend (Figure 16F).

Phylogeny

We used a conserved set of 192 coding sequences for phylogenetic construction. These sequences had >75% amino acid identity and 80% representation among the 20 species. Phylogenetic construction was performed using the GTRGAMMA model in RAxML (Stamatakis 2014). All bipartition had 100% bootstrap support based on 1000 repetitions. All orders, families, and genera formed monophyletic groups (Figure 17). Species from the 'complex' and 'robust' coral clades (Romano and Palumbi 1996; Kitahara et al. 2010) also formed monophyletic groups. Repetitions of tree building with less conserved sets of orthologs (70%, 60%, 50% and 40% representation among species) all produced the same topology, but with lower bootstrap values. For the species in which they overlapped, our tree agreed fully with that published by Kitchen et al. (2015).

Strongly methylated genes evolve slowly

Pairwise comparisons between orthologs from *A. millepora* and each other species revealed that strongly methylated genes evolve slowly. The trend was strongest for nonsynonymous substitutions (dN). When orthologs from *A. millepora* were compared with other *Acropora* species, mean dN was between 43% and 68% higher for weakly methylated genes than strongly methylated genes (Figure 17 and Figure 18). Pairwise comparisons with all species outside of the *Acropora* genus produced similar results, with mean dN between 17% and 52% (mean = 36 ± SEM 3%) higher for weakly methylated genes (Figure 17 and Figure 19). Negative correlation between dN and MBD-score was significant for all species comparisons ($p << 0.001$; Spearman's Rank Test).

The relationship between MBD-score and synonymous substitution rate (dS) was less pronounced than for dN, and varied with evolutionary proximity between species. Comparison of orthologs between *A. millepora* and other *Acropora* species showed no relationship (Figure 17).

Comparisons with corals outside of *Acropora* however, showed a significant negative relationship, with an average of 17% higher mean dS for weakly methylated genes (Figure 17 and Figure 20). The correlations with the three anemone species were weaker, although still significant. As most of these comparisons were saturated for synonymous substitutions they should be treated with caution. Analysis of dN/dS values gave similar results to dN for all groups of species (Figure 17).

Strongly methylated genes show greater codon bias

Because DNA methylation alters mutation patterns, we hypothesized that GBM shapes synonymous codon usage in stony corals. Specifically, we predicted that strong GBM produces codon bias via mutational replacement of codons bearing CpG dinucleotides (Kanaya et al. 2001; Qin et al. 2013). To test this we correlated MBD-scores with three distinct indices of codon bias: frequency of optimal codons (Fop)(Ikemura 1981), codon adaptation index (CAI)(Sharp and Li 1987a), and effective number of codons (Nc)(Wright 1990). Fop and CAI each quantify the preference for a set of optimal codons in the coding sequence. Higher values for these metrics indicate stronger codon bias. Nc quantifies nonrandom synonymous codon usage without assuming optimal codons. It is bounded between 1 (indicating complete bias, or use of only 20 codons for the 20 amino acids) and 64 (indicating completely neutral codon usage)(Wright 1990). All three indices correlated significantly with MBD-score (Figure 21). To assess the extent to which codon bias was driven by CpG hypermutability we recalculated CAI estimates using the same relative adaptiveness values (*W*)(see methods) for each codon, but excluding the five amino acids coded for by CpG bearing codons (Serine, Proline, Threonine, Alanine and Arginine). This substantially weakened the correlation from 0.38 (Figure 22A; Spearman's rho; $p << 0.0001$) to 0.17 (Figure 22B), although it remained significant ($p << 0.0001$). In contrast, recalculation of CAI based solely on these five amino acids strengthened the correlation (rho = 0.42; Figure 22C), indicating that hypermutability of CpGs due to GBM has a strong influence on codon usage.

CpG codons are under-represented in highly expressed genes

To further explore the influence of 5mC hypermutability on codon bias, we examined usage of CpG codons in highly expressed genes. As we did not have gene expression data for all species we first examined usage in annotated ribosomal protein genes with the assumption that these genes are highly expressed. For each species, relative synonymous codon usage (RSCU) of CpG codons was depressed in ribosomal protein genes (Figure 23A). To ensure that this did not result from variation in overall GC content we showed that mean RSCU of CpG codons was significantly lower than that of codons with GC, GG, or CC dinucleotides (t-tests; p for all species < 0.01).

For *A. millepora*, we assessed depression of CpG codons in highly expressed genes using three additional metrics: ΔRSCU (the difference in relative usage between the top 5% and bottom 5% expressed genes), rRSCU (the relative synonymous codon usage calculated for a concatenation of all ribosomal protein genes), and *W* (the relative adaptiveness of each codon; see methods). With one exception that had neutral usage, all CpG codons were underrepresented for all three metrics (Figure 23B-D). Hence CpG bearing codons are depressed in highly expressed genes.

Underrepresentation of CpG codons matches expectations for 5mC hypermutability

To further illustrate that loss of CpG codons is due to 5mC hypermutability, we examined RSCU for the five amino acids coded for by CpG bearing codons. Four of these, (Threonine, Proline, Alanine and Serine), are coded for by NCG codons, in which the CpG occupies the second and third positions of the codon. For these codons, 5mC>T mutations on the sense strand necessarily produce amino acid changes, which are expected to be rare due to purifying selection. In contrast, 5mC>T substitutions on the antisense strand produce silent substitutions (G>A within the codon)(Figure 24A). For this reason, we predicted that 5mC hypermutability would increase the usage of NCA codons at the expense of NCG codons. To show this, we plotted RSCU of synonymous codons against MBD-score, illustrating positive relationships for NCA codon usage (Spearman's rho between 0.156 and 0.196; $p << 0.001$) and opposing negative relationships for

NCG codon usage (Figure 25). Correlations of NCA codon usage with MBD-score were significantly stronger than other non-CpG codons (t-test; $p < 0.01$), indicating that NCA codons increase preferentially with stronger methylation. Hence for these four amino acids, depression of CpG codons in strongly methylated genes occurs through silent 5mC>T substitutions on the antisense strand. Moreover, all NCA codons were identified as optimal codons (Table 2), and their mean relative adaptiveness (for which the maximum is 1) was 0.99 (Table 3). These data indicate that that NCA codons replace NCG codons in strongly methylated genes.

The second group of CpG bearing codons is the CGN codons, which code for arginine. These are expected to evolve differently because 5mC>T substitutions on both the sense and antisense stands produce amino acid changes (Figure 24A). Although the trend is weak ($r = -0.06$; $p < 0.0001$), arginine content is negatively correlated with MBD-score (Figure 24B), suggesting a slight shift in arginine content due to CpG hypermutability.

Summarizing interrelationships between gene characteristics

To summarize the relationships between GBM and other gene characteristics we performed principal component analysis (PCA) on all coding regions for which we had MBD-scores and substitution rate estimates. Pair-wise estimates of dN and dS between *A. millepora* and *Siderastrea siderea* were used because it was the species outside of the genus *Acropora* with the greatest number of orthologs. Substitution rates based on other species produced qualitatively similar results. Variation in measures of GBM and codon bias was captured largely by the first principal component (34.0% variance explained)(Figure 26). While the indices of codon bias often correlated most strongly with one another, the strongest alternative predictor for all three was historical germ-line methylation as measured by CpGo/e (Table 4). Variation in transcript abundance, gene length, and substitution rates was captured largely by the second principal component (14.2% variance explained; (Figure 26).

DISCUSSION

Gene body methylation is a signature of broad and stable expression

We showed that strongly methylated genes in *A. millepora* tend to have constitutive and ubiquitous functions and are less likely to be differentially expressed across developmental stages and environmental regimes. These results corroborate earlier findings from diverse taxa including plants (Aceituno et al. 2008; Coleman-Derr and Zilberman 2012; Takuno and Gaut 2012), cnidarians (Sarda et al. 2012; Dimond and Roberts 2015), mollusk (Gavery and Roberts 2010), arthropods (Elango et al. 2009; Wang et al. 2013), and a basal chordate (Suzuki et al. 2013)(Keller et al. 2015). The relationship with differential expression in response to environmental regimes suggests the intriguing possibility that GBM could modulate gene expression plasticity.

We also found a positive correlation between GBM and transcript abundance, indicating that intermediately methylated genes are highly transcribed, while lowly methylated and extremely strongly methylated genes tend toward lower transcription. These results are similar to previous findings in plants (Zhang et al. 2006; Zilberman et al. 2007; Zemach et al. 2010), corals (Dimond and Roberts 2016), mollusk ((Gavery and Roberts 2013; Wang et al. 2014), and human (Jjingo et al. 2012), indicating that this connection between GBM and expression is evolutionarily ancient and widely conserved.

Gene body methylation and evolutionary rates

We show that GBM negatively correlates with substitution rates. This finding is consistent with previous results from pants (Takuno and Gaut 2012; Wang et al. 2015) and animals (Park et al. 2011; Sarda et al. 2012; Keller et al. 2015). Still, principal component analysis revealed that while substitution rates are significantly negatively correlated with GBM, they correlate more strongly with transcript abundance—a well-known trend described in bacteria, plants, fungi, and animals (Pál et al. 2001; Subramanian and Kumar 2004; Drummond et al. 2005; Drummond and Wilke 2008; Yang and Gaut 2011). This ubiquitous negative correlation between substitution rate

and expression is explained by stronger purifying selection against protein misfolding in highly expressed genes. Because they have a greater cumulative opportunity for errors and misfolding, mutations in highly expressed genes pose greater fitness costs than those in lowly expressed genes (Drummond et al. 2005). Similar logic can be applied to broadly expressed genes, as they are active in a greater number of cells and tissues (Duret and Mouchiroud 2000), must operate in a greater variety of cellular millieus (Hastings 1996), and undergo more translational events at the scale of the entire organism. Whereas non-synonymous substitutions affect the probability of protein misfolding through direct destabilization, synonymous substitutions most likely exert a similar but weaker effect by lowering translational accuracy (Akashi 1994; Drummond and Wilke 2008). Hence both dN and dS are expected to be lower in highly and broadly expressed genes. We have shown that in our system, highly expressed genes tend to be strongly methylated (Figure 16), and strongly methylated genes tend toward broad, constitutive transcription (Figure 14 and Figure 15). We conclude that the observed correlation between GBM and substitution rates is most parsimoniously explained by the occurrence of GBM on genes that are under stronger purifying selection because of their expression patterns. A corollary of this conclusion is that purifying selection generally outweighs the effects of 5mC hypermutability. Hence the paradox that GBM causes hypermutability but also correlates with sequence conservation can be explained by the fact that strongly methylated genes tend to undergo strong selection.

Gene body methylation shapes codon usage

Codon bias occurs for two reasons. The first is mutational bias, where differences in mutation rates across species and genomic contexts produce non-random variation in synonymous codon usage (Plotkin and Kudla 2011). The second mechanism is natural selection, which requires that synonymous mutations affect organismal fitness (Behura and Severson 2013). In our case, mutational processes mediated by GBM appear to be the stronger source of variation. We found that GBM correlates strongly with three separate indices of codon bias (Figure 21). Analysis of

RSCU values for NCG codons was consistent with codon bias arising largely through silent 5mC>T substitutions on the antisense stand (Figure 25). In other words, GBM causes codon bias by shifting usage of NCG codons to NCA codons.

When assessing whether codon bias is due to selection, researchers examine whether it occurs in genes where translation accuracy and efficiency are most important. Evidence that codon bias is due to selection includes: 1) positive correlation with expression level, 2) positive correlation with breadth of expression, and 3) negative correlation with synonymous substitution rate (Sharp and Li 1987b; Duret 2002; Zhang and Li 2004; Plotkin and Kudla 2011; Behura and Severson 2013). As ours and previous studies have shown, GBM covaries with each of these factors. In other words, GBM occurs on the types of genes predicted to undergo strongest selection on codon usage. This fact highlights the need for caution when attributing codon bias to selection, since in our case codon bias results largely from mutation. Here relationships with dS are of particular interest, because low dS can reflect selection on synonymous codons (Akashi 1994; Drummond and Wilke 2008). In our PCA, dS was nearly orthogonal to measures of GBM and codon bias. This result indicates that if *A. millepora* harbors codon bias due to selection, it is probably best predicted by expression level, and is dwarfed by mutational effect of GBM.

Although we attribute codon bias largely to mutation, this may still produce a potentially adaptive result—establishing a set of preferred and unpreferred codons in constitutively active genes. Optimal translation dynamics could then be achieved through evolution of tRNA abundances to match these preferred and unpreferred codons, obviating the need for selection of individual codons on a site-by-site basis. To put it another way, selection coefficients for individual synonymous codons will be exceedingly small (Bulmer 1987). In contrast, if a set of preferred codons is mutationally established in constitutively expressed genes, alleles that control the abundance of appropriate tRNAs could have stronger effects more amenable to natural selection. To be clear, we are not proposing that GBM originally evolved for this purpose. However, if its original function was linked with constitutively active expression, as appears to be the case from studies of plants (Takuno and Gaut 2012), invertebrates (Sarda et al. 2012) and mammals (Baubec

et al. 2015), then CpG replacement coupled with coevolution of tRNAs provides an efficient means of evolving optimal codons in the genes where they are most beneficial.

An advantage of mutation-driven codon bias is that it could be maintained even in the absence of efficient selection, so it would be particularly beneficial for organisms with relatively small population size or otherwise inefficient selection. If some adaptive value of GBM is indeed related to maintenance of codon usage, it is not surprising that organisms such as yeast, fly and worm are able to exist without it (Capuano et al. 2014); due to their large population sizes their optimal codon usage can be maintained by selection alone. At a minimum, it seems likely that tRNA pools have evolved in response to methylation-induced codon bias. This hypothesis could be explored through phylogenetic comparison of tRNA abundances between clades that independently lost or retained GBM.

Conclusions and outlook

Here we present three primary findings on gene body methylation in stony corals: 1) GBM is most pronounced in genes with broad and stable expression; 2) GBM predicts sequence conservation 3) hypermutability due to GBM drives codon bias. Conserved occurrence of GBM on constitutively expressed genes in plants and the basal metazoan examined here indicates an evolutionarily ancient function involving selective pressure for accurate and stable gene expression. One means of improving translation fidelity is the use of optimal codons. Given its capacity to establish preferred and unpreferred codons in actively expressed genes, GBM could potentially influence evolution of optimal codons.

Table 1 Sources of transcriptomic data

order	family	Genus	species	Citation
Actiniaria	Actiniidae	*Anthopleura*	*elegantissima*	Kitchen et al. 2015
Actiniaria	Aiptasiidae	*Aiptasia*	*pallida*	Lehnert et al. 2012
Actiniaria	Edwardsiidae	*Nematostella*	*vectensis*	Nordberg et al. 2014
Scleractinia	Acroporidae	*Acropora*	*cervicornis*	Libro et al. 2013
Scleractinia	Acroporidae	*Acropora*	*palmata*	Polato *et al.* 2011
Scleractinia	Acroporidae	*Acropora*	*hyacinthus*	Barshis et al. 2013
Scleractinia	Acroporidae	*Acropora*	*tenuis*	none
Scleractinia	Acroporidae	*Acropora*	*millepora*	Moya et al. 2012
Scleractinia	Acroporidae	*Acropora*	*digitifera*	Shinzato et al. 2011
Scleractinia	Astocoeniidae	*Madracis*	*auretenra*	none
Scleractinia	Faviidae	*Platygyra*	*carnosus*	Sun et al. 2013
Scleractinia	Faviidae	*Platygyra*	*daedalea*	none
Scleractinia	Fungiidae	*Fungia*	*scutaria*	Kitchen et al. 2015
Scleractinia	Merulinidae	*Orbicella*	*faveolata*	Anderson et al. 2016
Scleractinia	Montastraeidae	*Montastraea*	*cavernosa*	Kitchen et al. 2015
Scleractinia	Mussidae	*Pseudodiploria*	*strigosa*	none
Scleractinia	Pocilloporidae	*Pocillopora*	*damicornis*	Traylor-Knowles et al. 2011
Scleractinia	Pocilloporidae	*Seriatopora*	*hystrix*	Kitchen et al. 2015
Scleractinia	Pocilloporidae	*Stylophora*	*pistillata*	Maor-Landaw et al. 2014
Scleractinia	Poritidae	*Porites*	*astreoides*	Kenkel et al. 2013
Scleractinia	Poritidae	*Porites*	*lobata*	none
Scleractinia	Poritidae	*Porites*	*australiensis*	Shinzato et al. 2014
Scleractinia	Siderastreidae	*Siderastrea*	*siderea*	Davies et al. 2016

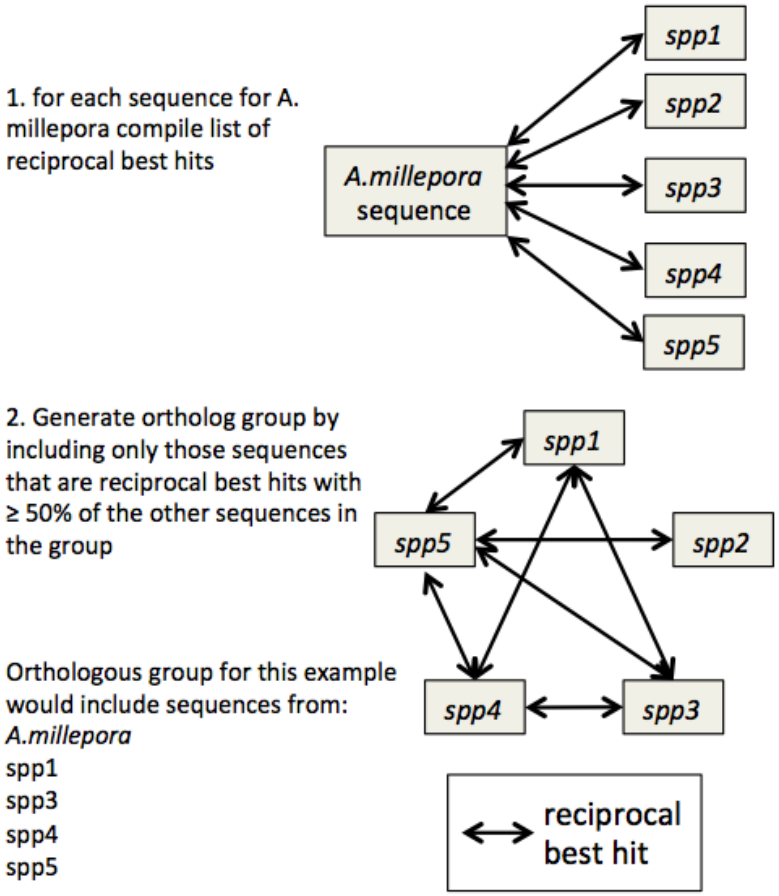

Figure 10 Schematic representation of ortholog assignment method. Sequences from *A. millepora* were used as anchors. For each sequence, reciprocal best hits from each other species were assembled as candidate orthologs. This group of candidates was then subset by iteratively removing sequences that were reciprocal best its with < 50% of other sequences within the group.

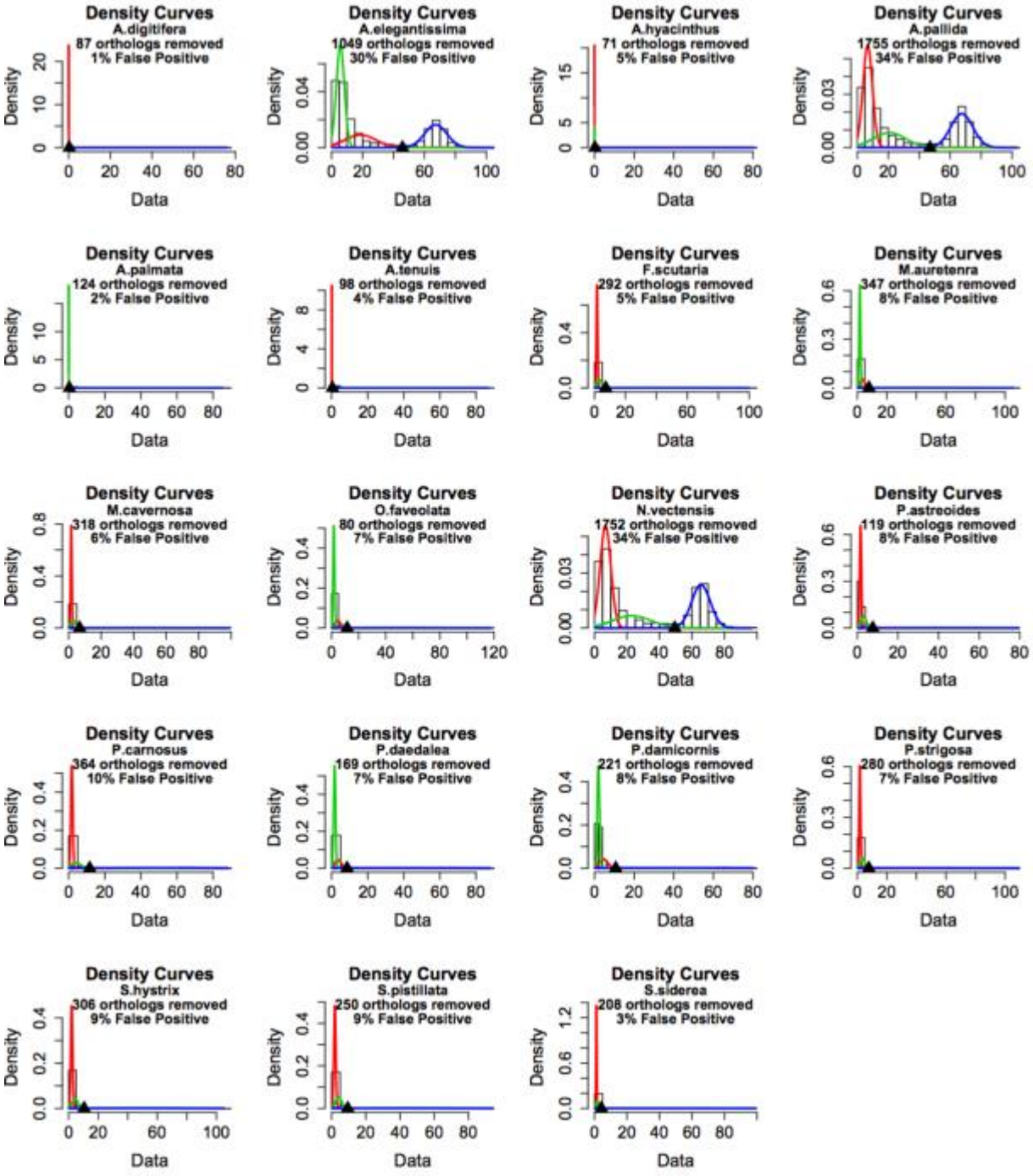

Figure 11 Identification and removal of false-positive ortholog calls. A three component Gaussian mixture model was fitted to the pairwise dS estimates with *A. millepora* for each species. The third component (blue above) was assumed to represent false positives. These orthologs (to the right of the black triangle) were removed from further analysis. The number and percentage of false positives removed is given in the title for each figure. The three anemone species, (*A. elegantissima*, *A. pallida*, and *N. vectensis*) displayed much greater rates of false positives.

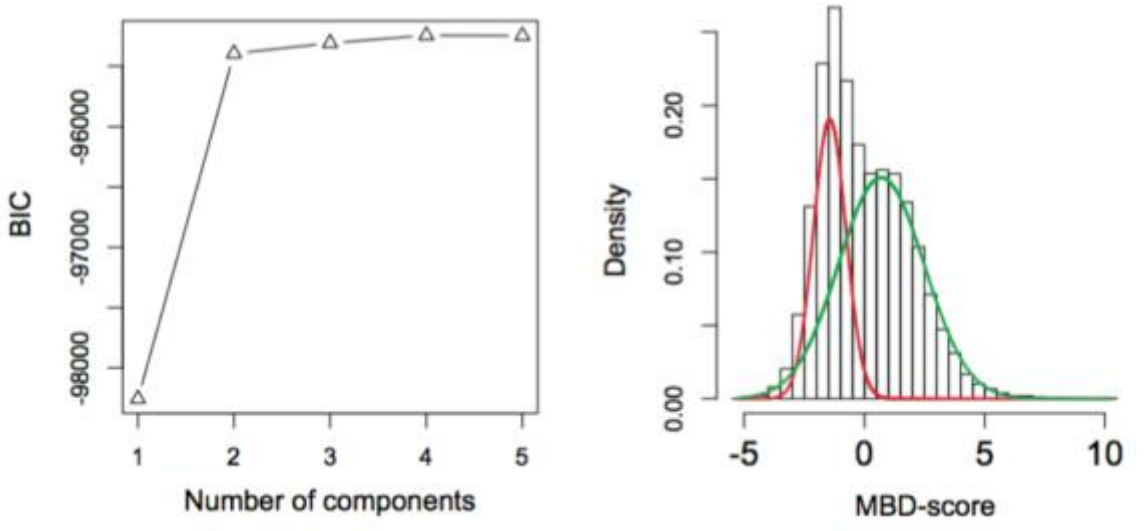

Figure 12 Fitting of Gaussian mixture components to distribution of MBD-scores. (A) Plot of Bayesian Information Criteria for models of the distribution of MBD-scores using different numbers of Gaussian components. (B) Traces of the two-component model overlaid on the distribution.

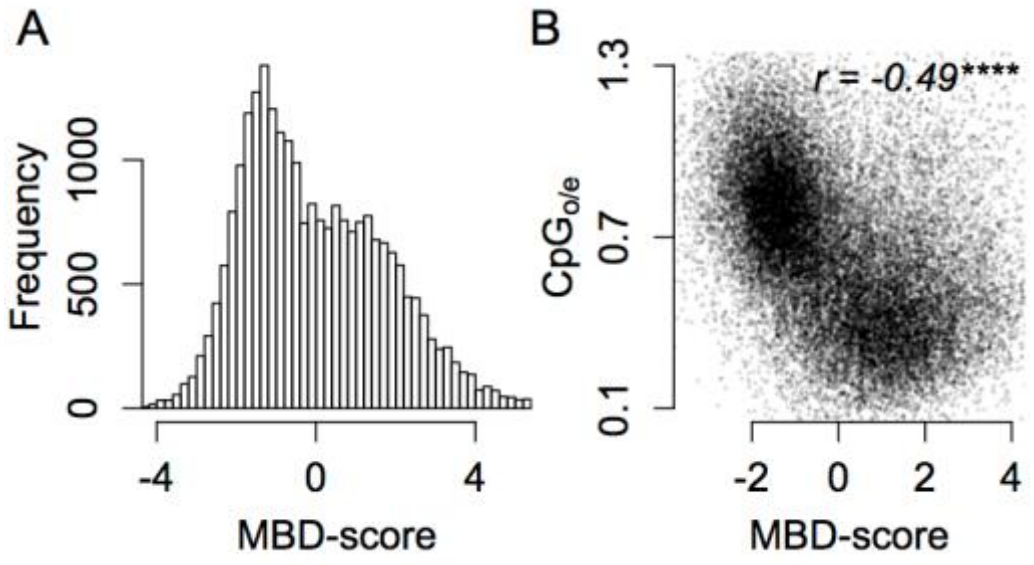

Figure 13 MBD-score is bimodally distributed and correlates with CpGo/e. (*A*) Distribution of MBD-score (log_2-fold difference between enriched and flow-through MBD-seq libraries). Higher values indicate stronger methylation. (*B*) Scatter plot of MBD-score and CpGo/e. Lower values for CpGo/e are expected with stronger methylation. Asterisks indicate significance based on Spearman's rank-order correlation test (ns >0.05; * <0.05; ** <0.01; *** <0.001; **** <0.0001).

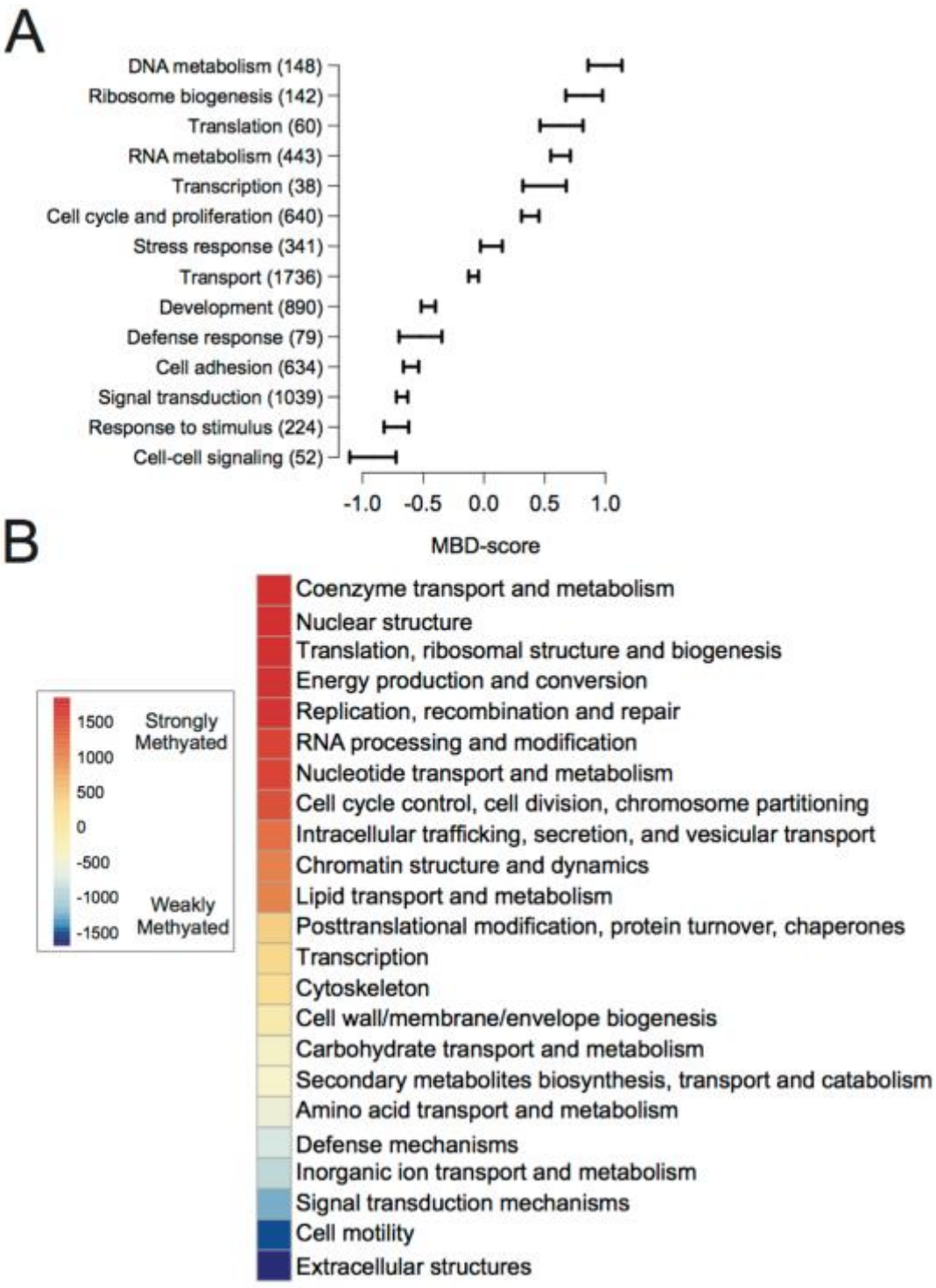

Figure 14 Relationship between gene functional categories and MBD-score. (A) Mean MBD-score for a selected set of Gene Ontology (GO) terms for biological processes. Error bars indicate standard error. (B) Enrichment of KOG terms based on Mann-Whitney U tests implemented in the R package KOGMWU as in Dixon et al. (2015).

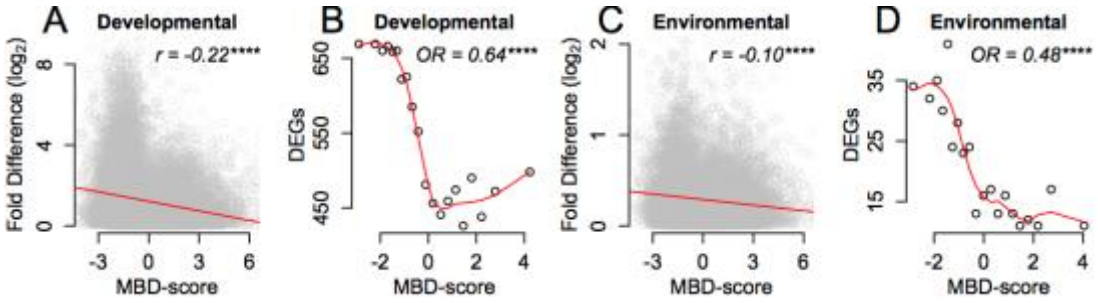

Figure 15 GBM predicts transcriptional stability across developmental stages and environmental regimes. (*A*) Scatter plot of MBD-score and transcriptional variation (given as $\log_2$-fold differences) between adult colonies and juvenile offspring. Red line shows least squared regression. Asterisks indicate significance based on Spearman's rho. (*B*) Distribution of differentially expressed genes (DEGs; FDR <0.01) between juveniles and adults. All genes were divided into 20 quantiles ranked by MBD-score. The number of differentially expressed genes in each quantile was plotted against the median MBD-score for that quantile. Enrichment of DEGs among the weakly methylated genes (MBD-score <0) compared with strongly methylated genes (MBD-score ≥0) is given as the odds ratio (OR) for Fisher's exact test. Red line shows a smoothed trace of the points. (*C*, *D*) The same figures representing transcriptional variation between populations of clonal colony fragments transplanted between distinct habitats described in Dixon et al. (2014). Asterisks indicate significance based on Spearman's rank-order correlation test (ns >0.05; * <0.05; ** <0.01; *** <0.001; **** <0.0001).

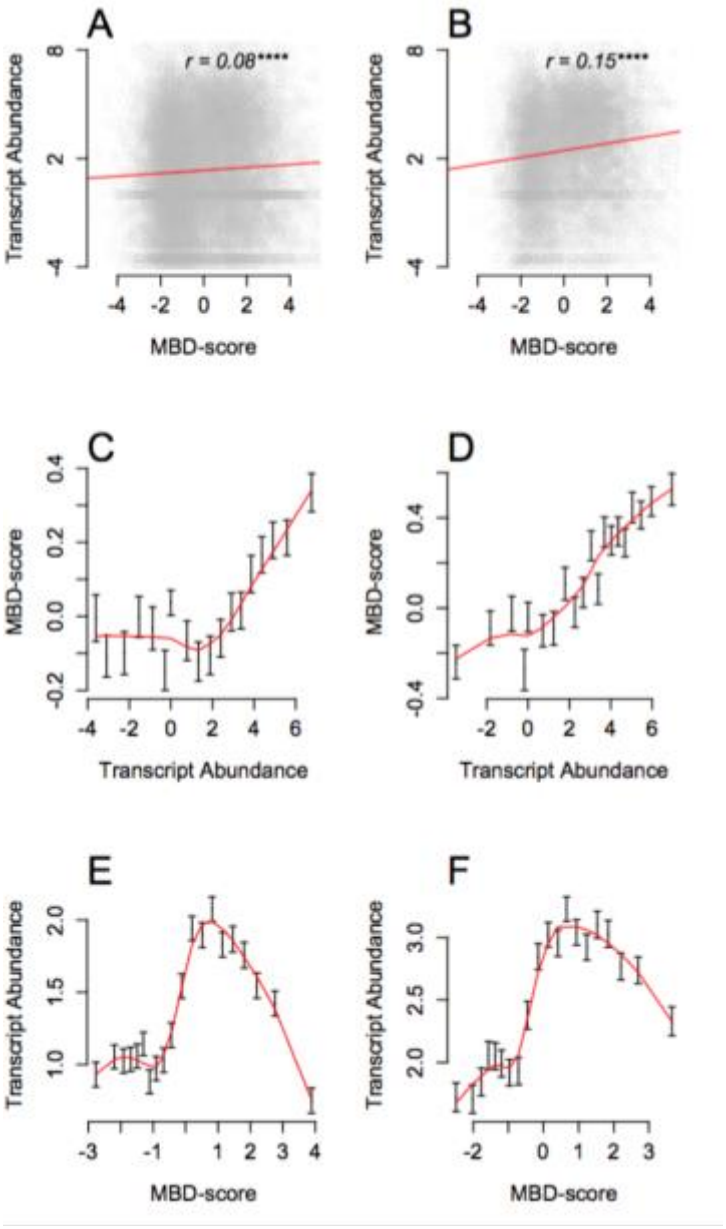

Figure 16 Relationships between transcript abundance and MBD-score. Figures are paired to illustrate interrelationship with gene length. Left panels show relationships for all coding sequences, right panels for coding sequences longer than 800 bp. (A-B) Correlation between MBD-score and normalized transcript abundance. Correlation is given as Spearman's Rho (r). Asterisks denote significance based on Spearman's rank tests. Red line traces least squared linear regression. (C-D) Highly expressed genes tend to be strongly methylated. Mean MBD-score was plotted for 12 quantiles of genes ranked by transcript abundance. Error bars indicate standard error. (E-F) MBD-score generally predicts higher expression, but the most strongly methylated genes show lower expression. This effect is especially true for shorter genes, an effect also described in *Arabidopsis* (Zilberman et al. 2007). Significance notation: ns > 0.05; * < 0.05; ** < 0.01; ***<0.001; **** < 0.0001.

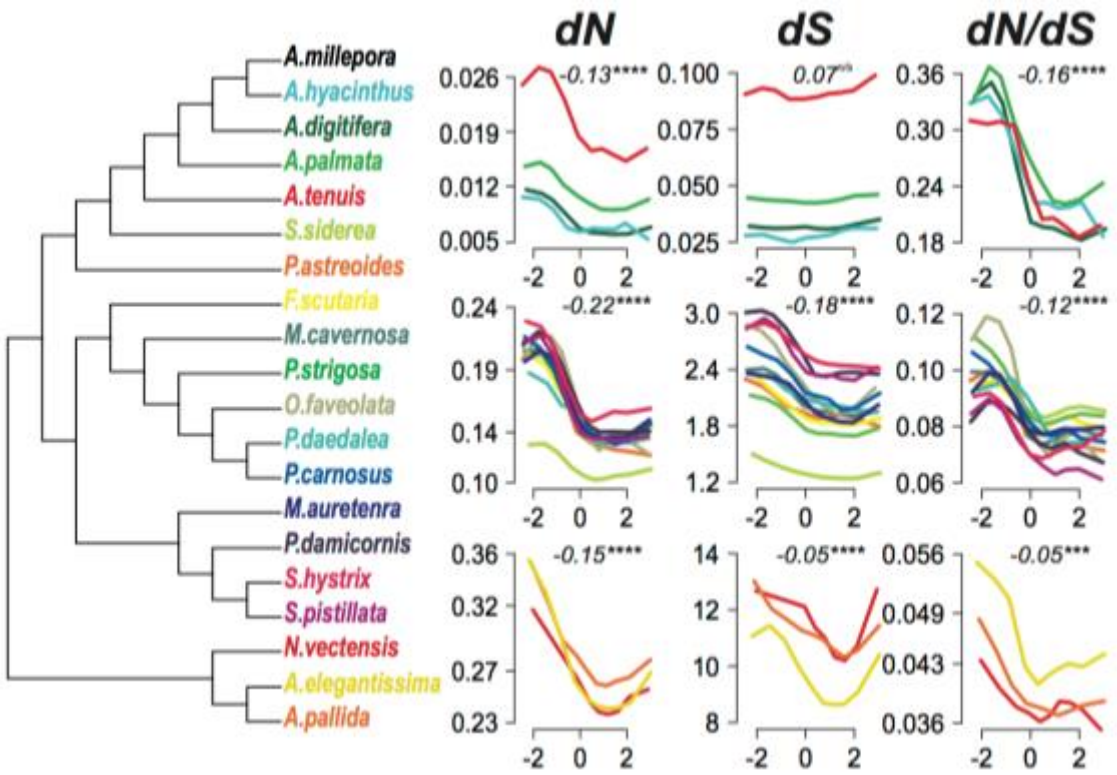

Figure 17 Relationship between MBD-score and substitution rates across the anthozoan phylogeny. All nodes in the phylogeny have 100% bootstrap support based on 1,000 replicates. Line plots trace the mean substitution rates for all genes divided into 10 quantiles ranked by MBD-score. Line color indicates which species *A. millepora* was compared with to estimate pair-wise substation rates. The top row of line plots shows comparisons within *Acropora*. The middle row shows corals outside of *Acropora*. The third row shows comparisons with anemone species. For each panel, the correlation (Spearman's rho) and statistical significance indicate the median values across all included species. Individual correlations are reported in Figure 18 and Figure 19. Asterisks indicate significance based on Spearman's rank-order correlation test (ns >0.05; * <0.05; ** <0.01; *** <0.001; **** <0.0001).

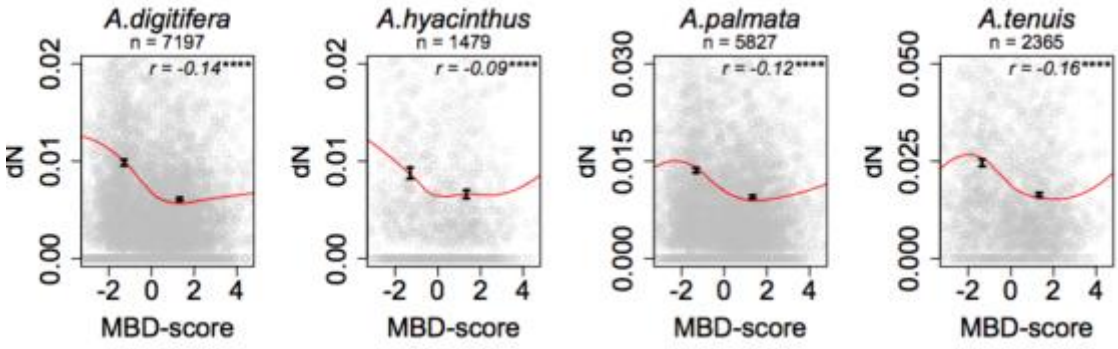

Figure 18 Relationship between nonsynonymous substitution rate (dN) and MBD-score across all species outside of *Acropora*. The two error bars in each panel display mean dN and standard error for the strongly methylated (MBD-score >= 0) and weakly methylated (MBD-score < 0) genes. Correlations are given as Spearman's Rho. All p values for Spearman's rank correlation test were < 0.0001. Red lines are smoothed traces with a span of 0.8.

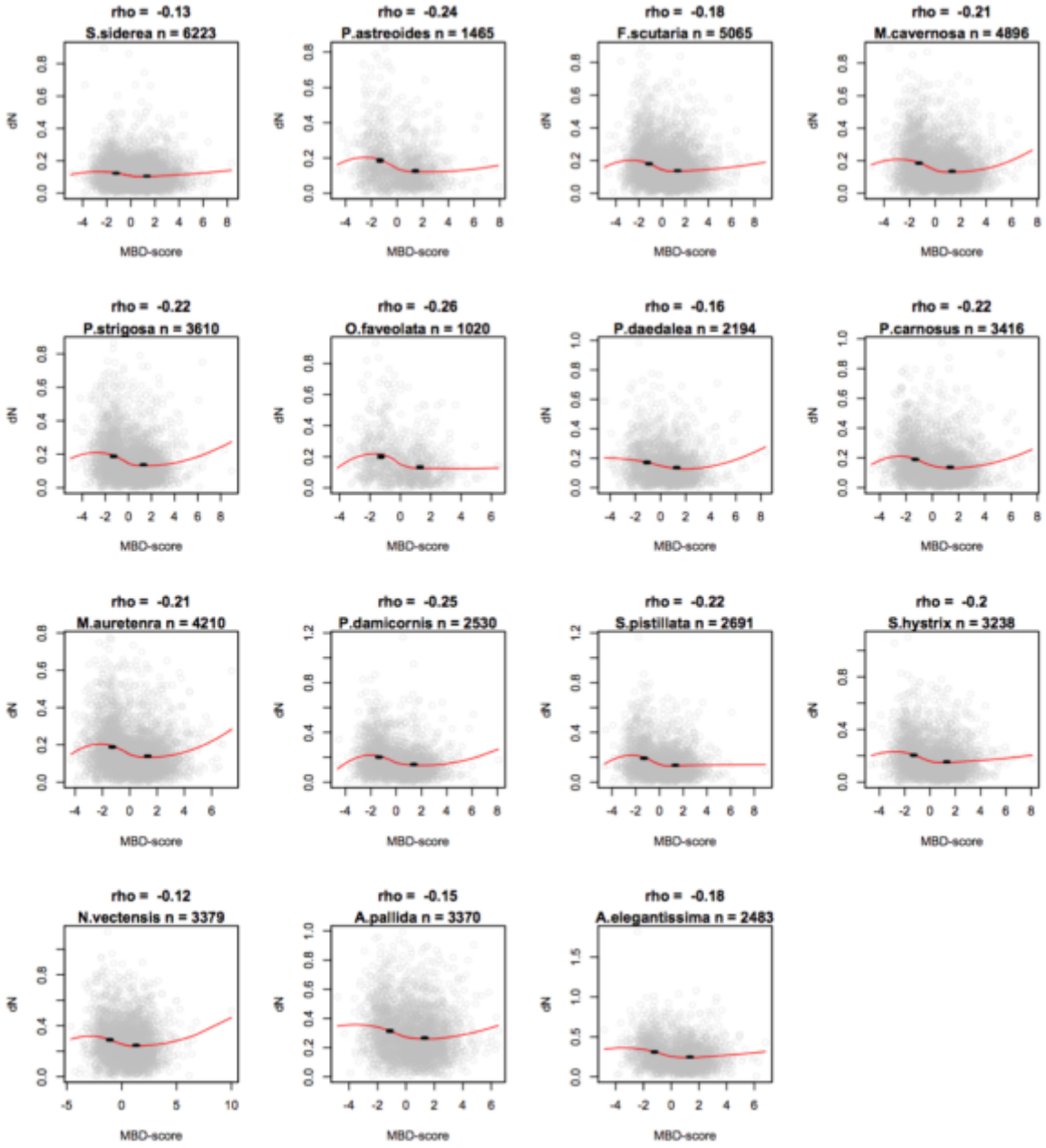

Figure 19 Relationship between nonsynonymous substitution rate (dN) and MBD-score across all species outside of *Acropora*. The two error bars in each panel display mean dN and standard error for the strongly methylated (MBD-score >= 0) and weakly methylated (MBD-score < 0) genes. Correlations are given as Spearman's Rho. All p values for Spearman's rank correlation test were < 0.0001. Red lines are smoothed traces with a span of 0.8.

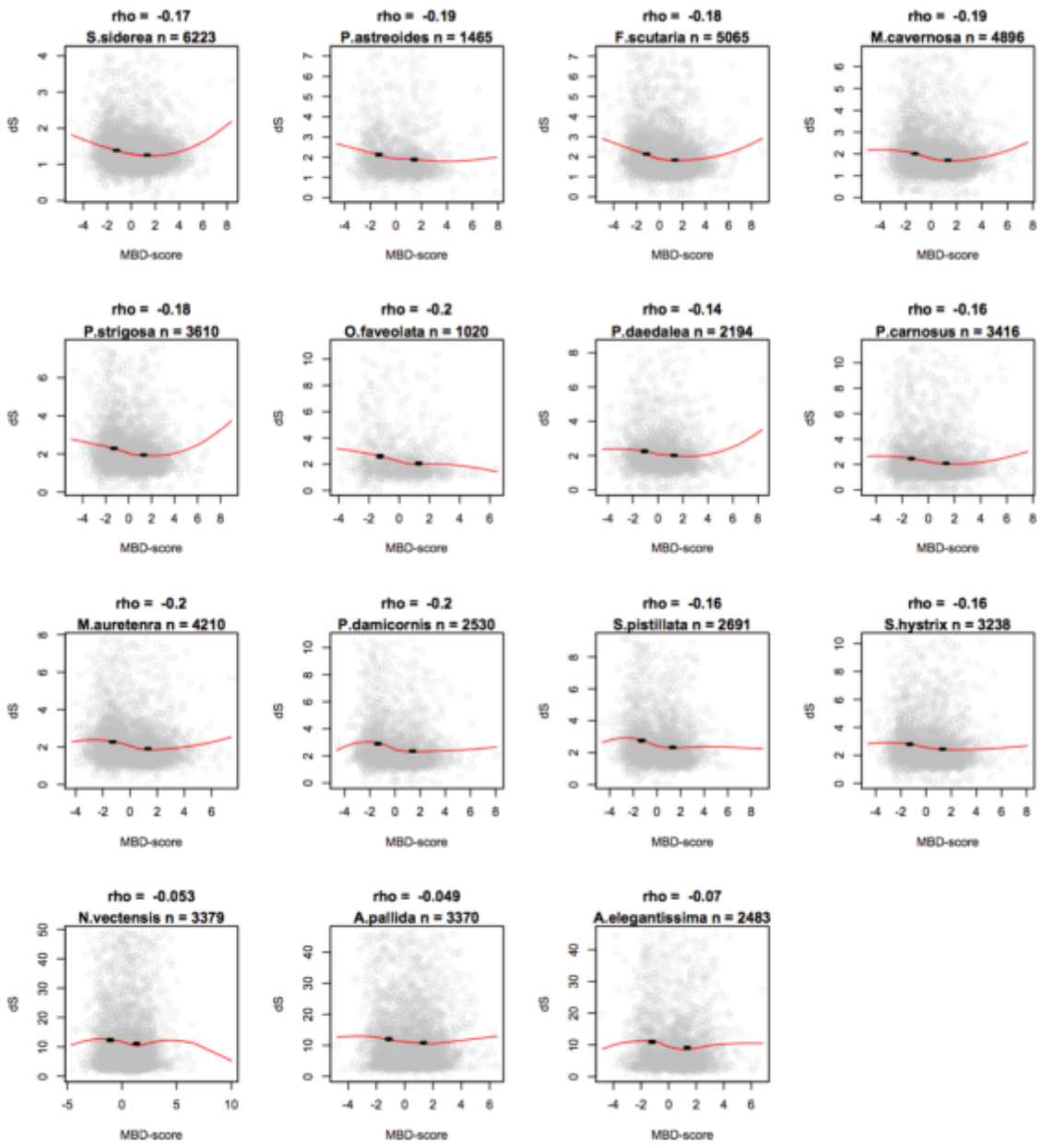

Figure 20 Relationship between synonymous substitution rate (dS) and MBD-score across all species outside of *Acropora*. The two error bars in each figure display mean dS and standard error for the strongly methylated (MBD-score >= 0) and weakly methylated (MBD-score < 0) genes. Correlation is given as Spearman's Rho. All p values for Spearman's rank correlation test were < 0.0001. Red lines are smoothed traces with a span of 0.8.

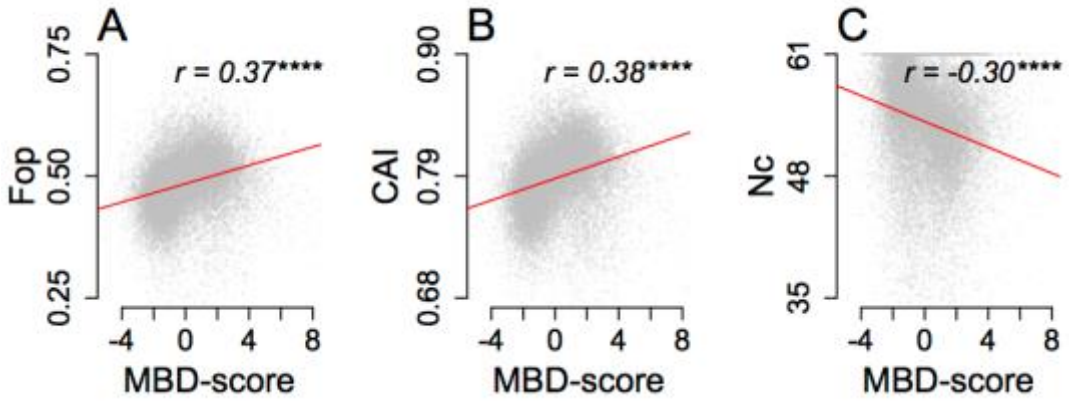

Figure 21 Correlation between MBD-score and indices of codon bias. (*A*) Fop. (*B*) CAI. (*C*) Nc. Red lines trace least squared linear regression. Asterisks indicate significance based on Spearman's rank-order correlation test (ns >0.05; * <0.05; ** <0.01; *** <0.001; **** <0.0001).

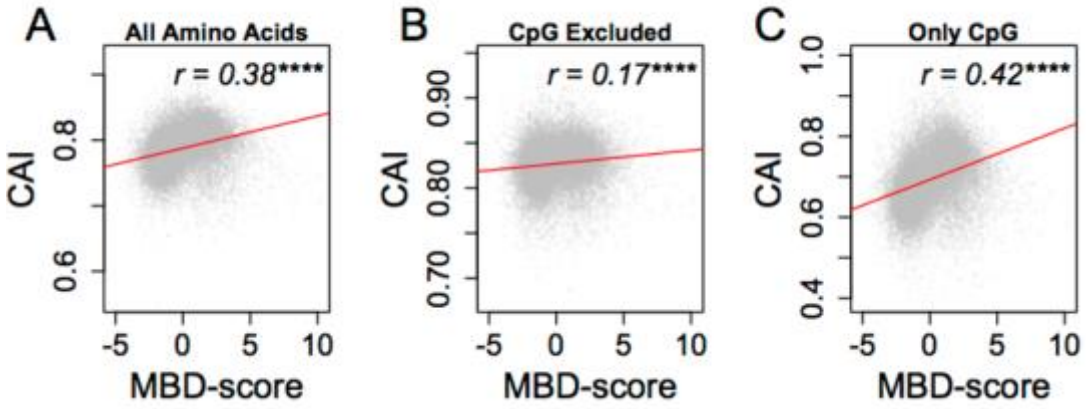

Figure 22 Correlation between codon adaptation index (CAI) and MBD-score with and without amino acids coded for by codons with CpG dinucleotides (Serine, Proline, Threonine, Alanine and Arginine). (A) Correlation between CAI and MBD-score with all amino acids included. (B) Calculating CAI with Serine, Proline, Threonine, Alanine and Arginine severely reduces correlation. (C) Calculating CAI based solely on Serine, Proline, Threonine, Alanine and Arginine increases strengthens correlation. Asterisks indicate significance based on Spearman's rank tests. Red lines trace least squared regression.

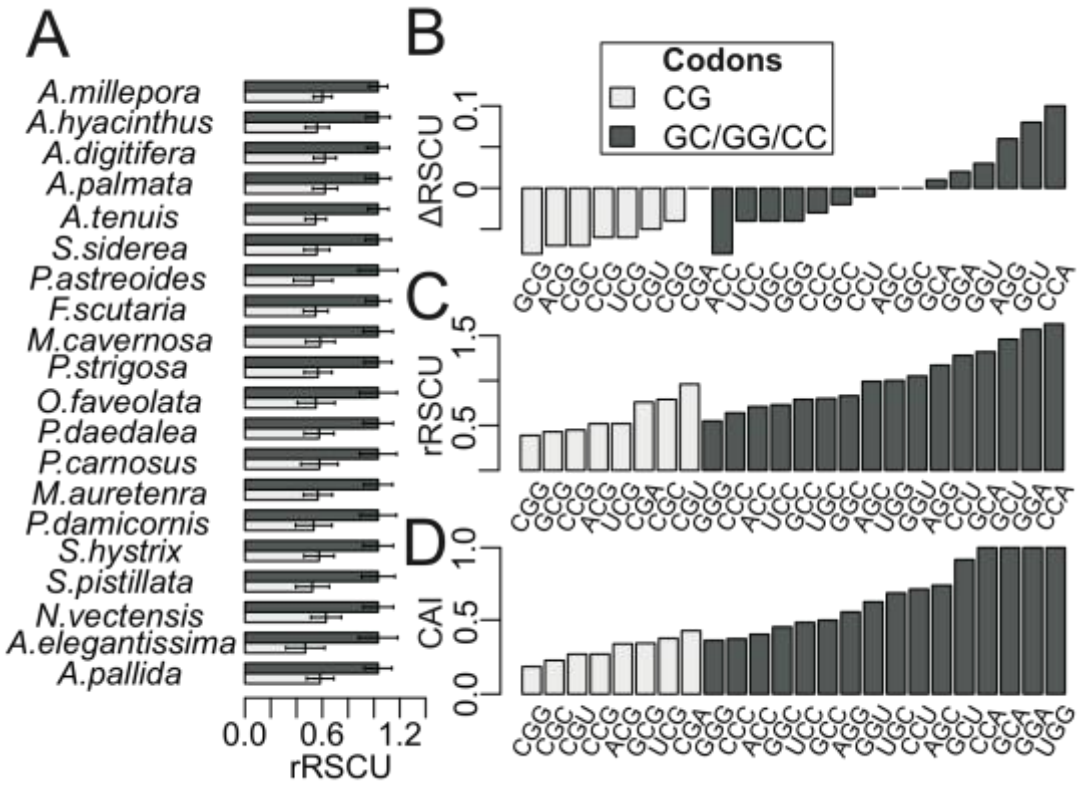

Figure 23 Codons bearing CpG dinucleotides are underrepresented in highly expressed genes. A) Comparison of mean Relative Synonymous Codon Usage for CG bearing codons compared to all other codons bearing GC, GG, or CC dinucleotides in ribosomal genes. Error bars show standard error. A value of 1 for this metric indicates no bias. B) Comparison of ΔRSCU for codons bearing CG, GC, GG or CC dinucleotides in *A. millepora*. ΔRSCU is the difference in RSCU between the top 5% most highly expressed genes and bottom 5%. Negative values indicate underrepresentation in highly expressed genes. C) Comparison of RSCU for CG, GC, GG, or CC codons in ribosomal genes from *A. millepora*. Values less than one indicate underrepresentation in ribosomal genes. D) Comparison of relative adaptiveness (*W*) for CG, GC, GG, or CC codons in *A. millepora*. Here a value of 1 indicates that the codon is optimal for its amino acid. No CpG codons were optimal, and were all less than half as frequent as the optimal codon.

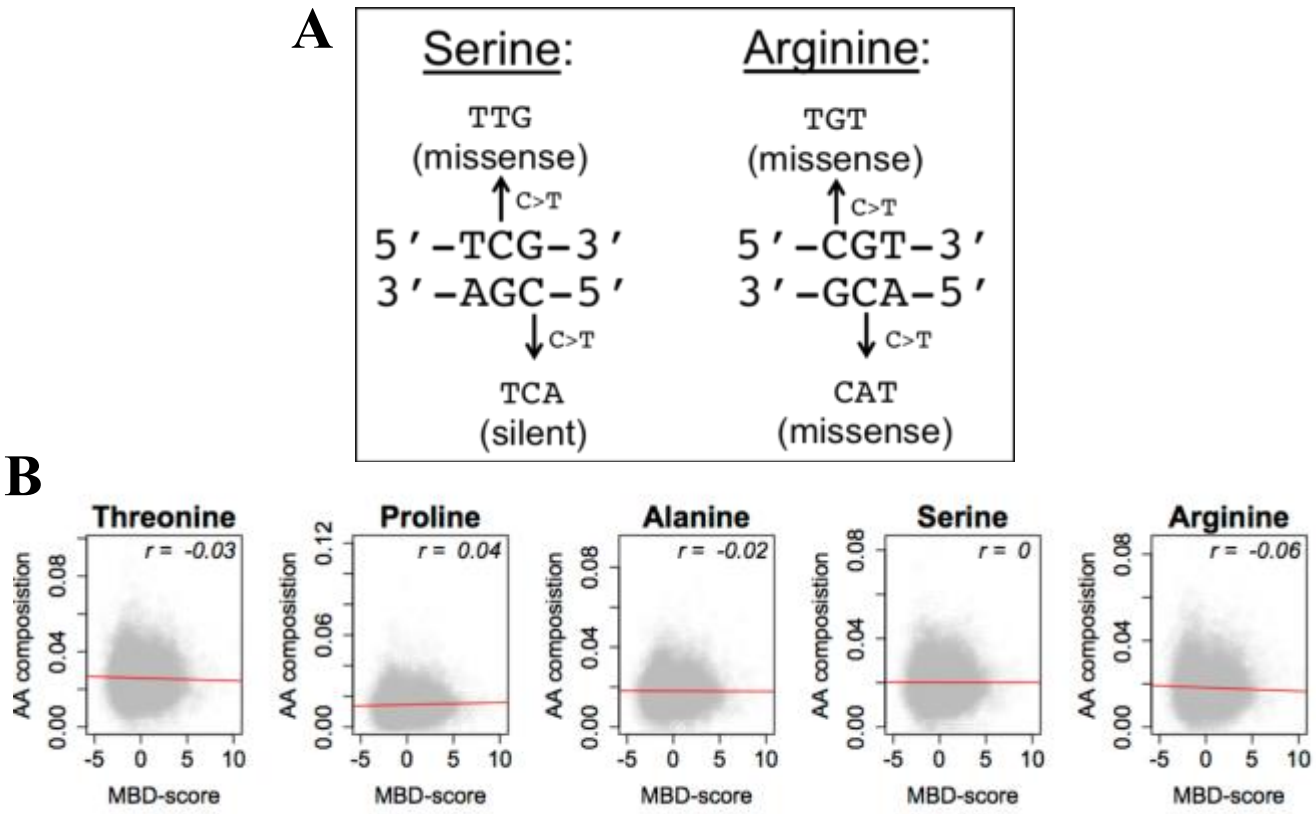

Figure 24 Loss of CpG bearing codons occurs through silent C>T substitution on the antisense stand. Methylated cytosines tend to be substituted for thymine (Shen et al. 1994). (A) On the sense strand, 5mC>T substitutions result in amino acid changes, whereas 5mC>T substitutions on the antisense strand are silent. (B) MBD-score shows little correlation with amino acid content, indicating that purifying selection counteracts most nonsynonymous 5mC>T substitutions. Although the correlation is weak, arginine content shows a stronger negative correlation than of the other amino acid. This is consistent with the fact for CGN codons, 5mC>T substitutions on either strand will replace the arginine.

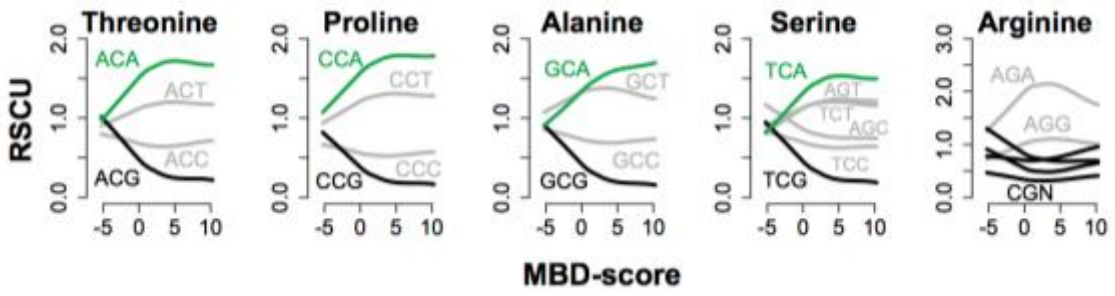

Figure 25 Depression of CpG bearing codons occurs via replacement with synonymous NCA codons. Lines show smoothed traces of the relationship between RSCU and MBD-score for the indicated codon. Black lines indicate CpG bearing codons. Green lines indicate NCA codons. Grey lines indicate all other codons. Opposing trends for NCA and NCG codons support the inference that NCA codons replace NCG codons in strongly methylated genes.

Table 2 Optimal codons identified based on correspondence analysis of codon usage implemented in CodonW. X^2 indicates the chi-square statistic describing the enrichment of the synonymous codons (see supplemental methods below). All NCA codons were identified as optimal and tended toward higher X^2. No NCG codons were optimal.

codon	amino acid	X^2
AGA	Arg	562.906
UUU	Phe	389.064
ACA	Thr	309.689
GCA	Ala	286.519
AGG	Arg	277.184
UCA	Ser	188.606
AAU	Asn	143.76
AGU	Ser	132.991
CCA	Pro	132.423
UAU	Tyr	111.065
AUA	Ile	95.299
GAU	Asp	91.223
CAU	His	85.921
UGU	Cys	59.489
UUA	Leu	49.41
GGA	Gly	48.125
AUU	Ile	37.656
GUG	Val	31.472
CUA	Leu	27.806
CUG	Leu	19.063
GGG	Gly	18.105
GUA	Val	17.354
AAG	Lys	8.718
GAA	Glu	4.369

Table 3 Relative adaptiveness of codons in *Acropora millepora* (see methods). NCA codons are highlighted in green and tend to have values equal to close to the maximum of 1.00. NCG codons are highlighted in red and always have the lowest relative adaptiveness value for their respective amino acids.

codon	amino acid	wi	rscu	codon	amino acid	wi	rscu
GCU	Ala	1.00	1.39	UUG	Leu	1.00	1.45
GCA	Ala	0.96	1.33	CUU	Leu	0.97	1.40
GCC	Ala	0.58	0.80	CUG	Leu	0.79	1.15
GCG	Ala	0.34	0.47	UUA	Leu	0.57	0.82
AGA	Arg	1.00	1.98	CUC	Leu	0.48	0.69
AGG	Arg	0.57	1.13	CUA	Leu	0.34	0.50
CGA	Arg	0.48	0.96	AAA	Lys	1.00	1.09
CGU	Arg	0.41	0.82	AAG	Lys	0.83	0.91
CGC	Arg	0.34	0.67	AUG	Met	1.00	1.00
CGG	Arg	0.22	0.44	UUU	Phe	1.00	1.22
AAU	Asn	1.00	1.09	UUC	Phe	0.64	0.78
AAC	Asn	0.83	0.91	CCA	Pro	1.00	1.63
GAU	Asp	1.00	1.21	CCU	Pro	0.78	1.27
GAC	Asp	0.65	0.79	CCC	Pro	0.39	0.64
UGU	Cys	1.00	1.14	CCG	Pro	0.28	0.46
UGC	Cys	0.75	0.86	UCA	Ser	1.00	1.36
CAA	Gln	1.00	1.06	AGU	Ser	0.88	1.20
CAG	Gln	0.89	0.94	UCU	Ser	0.88	1.20
GAA	Glu	1.00	1.21	AGC	Ser	0.70	0.95
GAG	Glu	0.65	0.79	UCC	Ser	0.56	0.76
GGA	Gly	1.00	1.56	UCG	Ser	0.39	0.53
GGU	Gly	0.71	1.11	ACA	Thr	1.00	1.53
GGC	Gly	0.52	0.81	ACU	Thr	0.78	1.19
GGG	Gly	0.34	0.53	ACC	Thr	0.49	0.75
CAU	His	1.00	1.14	ACG	Thr	0.35	0.53
CAC	His	0.75	0.86	UGG	Trp	1.00	1.00
AUU	Ile	1.00	1.4	UAU	Tyr	1.00	1.02
AUC	Ile	0.67	0.94	UAC	Tyr	0.96	0.98
AUA	Ile	0.48	0.67	GUU	Val	1.00	1.37
				GUG	Val	0.89	1.22
				GUC	Val	0.57	0.78
				GUA	Val	0.47	0.64

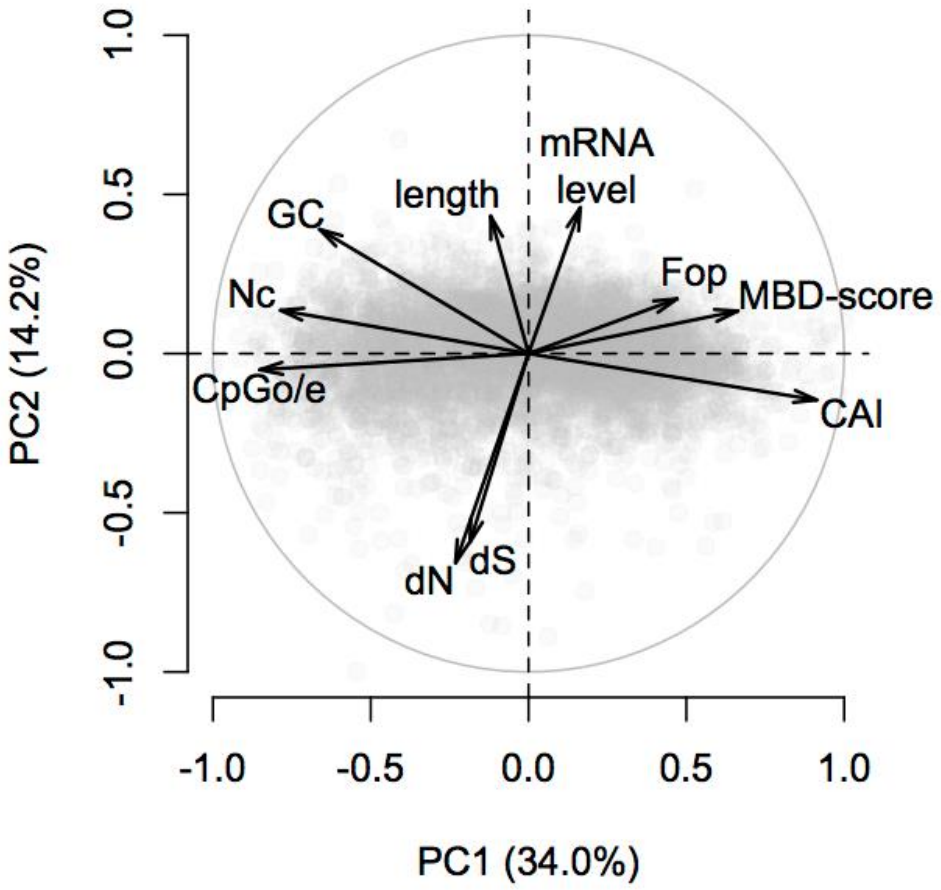

Figure 26 PCA of gene features in *A. millepora*. The first principal component explained 34.0% of variation and correlated primarily with measures of gbM and codon bias. The second principal component explained 14.2% of variation and correlated primarily with gene length, transcript abundance, and substitution rates. Variables included in the analyses are: normalized CpG content (CpGo/e), Nc, GC content of coding regions (GC), nonsynonymous substitution rate (dN), synonymous substitution rate (dS), length of coding region (length), transcript abundance (mRNA level), Fop, $\log_2$-fold difference between captured and flow-through fractions of methylation binding domain enrichment libraries (MBD-score), and CAI. Substitution rates are pair-wise estimates between *A. millepora* and *S. siderea*.

Table 4 Spearman's rank correlations between gene characteristics: Codon adaptation index (CAI), Effective number of codons (Nc), Frequency of optimal codons (Fop), log2 fold difference between methylation binding domain captured and flow-through fractions (MBD-score), transcript abundance (mRNA), length of the coding region (length), normalized CpG content (CpGo/e), and GC content (GC).

Variable	CAI	Nc	Fop	MBD-score	mRNA	length	CpGo/e	GC
CAI	1.00	-0.52	0.33	0.38	0.16	0.05	-0.71	-0.61
Nc		1.00	-0.24	-0.31	-0.02	0.17	0.44	0.34
Fop			1.00	0.17	0.18	0.09	-0.33	-0.01
MBD-score				1.00	0.08	0.05	-0.51	-0.22
mRNA					1.00	0.36	-0.12	0.00
length						1.00	-0.08	0.11
CpGo/e							1.00	0.33
GC								1.00

Chapter 3: On the role of gene body methylation in acclimatization

ABSTRACT

Because it can change through the lifespan of an organism and influence gene expression, DNA methylation thought to be a potential mediator of acclimatization. However, evidence for this hypothesis under ecologically relevant conditions remains scarce. Here we examined patterns of transcription and gene body methylation (GBM) in colony fragments of the coral *Acropora millepora* transplanted between distinct natural environments in the Great Barrier Reef to assess the hypothesis that GBM mediates acclimatization. Using discriminant analysis of principal components, we show that among coral transplants, resemblance in GBM patterns to native corals predicted elevated fitness. Environmentally driven variation in GBM did not however, correlate with transcription. Hence our results support the conclusion that patterns of GBM reflect functionally important genomic mechanisms, but do not appear to directly regulate transcript abundance.

INTRODUCTION

DNA methylation is a covalent chromatin modification that influences transcription in plants, animals, and fungi. The relative stability of this modification gives it unique potential as an adaptive mechanism. Whereas genetic adaptation must be sculpted by natural selection within populations, DNA methylation can change throughout individual life-histories (Law and Jacobsen 2010), and in response to environmental stimuli (Feil and Fraga 2011). Compared to transcription however, methylation is stable, and has much greater potential for transgenerational inheritance (Herman et al. 2016)(Wang et al. 2016). DNA methylation therefore represents a middle ground between the rigidity of genotype and the transience of gene expression. These characteristics are the basis for hypotheses that DNA methylation mediates phenotypic plasticity and facilitates adaptation (Angers et al. 2010)(Roberts and Gavery 2012)(Verhoeven et al. 2016)(Putnam et al.

2016)(Hofmann 2017). Evidence for these hypotheses in marine invertebrates however, remains scarce.

In this study, we investigate the role of DNA methylation in acclimatization. Our study system is the reef-building coral, *Acropora millepora*: a basal metazoan uniquely amenable to ecological epigenetics because individuals can be fragmented into genetically identical replicates. Understanding acclimatization in this system is also of special importance because of corals' high vulnerability to climate change (Foden et al. 2013). Using a reciprocal transplantation paradigm, we evaluate the role of a particular type of DNA methylation, gene body methylation, in coral acclimatization.

Gene body methylation (GBM) refers to methylation within the transcribed regions of coding genes, most often on cytosines with CG dinucleotides (CpGs)(Suzuki and Bird 2008). Although it occurs in both plants and animals (Zilberman et al. 2007)(Zemach et al. 2010), and may be the ancestral form of DNA methylation in Eukarya (Zemach et al. 2010), the adaptive function of GBM, if any, remains uncertain. GBM does however, demonstrate consistent associations with transcriptional patterns. In both plants and animals, GBM correlates with expression level and gene responsiveness (Zilberman 2017). Constitutively expressed genes (ie housekeeping genes) tend to be strongly methylated and inducible genes tend to be weakly methylated. This association extends to environmentally driven expression(Dixon et al. 2014)(Dimond and Roberts 2016)(Dixon et al. 2016), suggesting that GBM may be involved in modulating phenotypic plasticity (Roberts and Gavery 2012).

To better understand the role of GBM in phenotypic plasticity, we assayed genome-wide patterns of DNA methylation in coral fragments transplanted to different sites in the Great Barrier Reef. Thirty colonies of *A. millepora* were divided into fragments and reciprocally transplanted between a warmer site, Orpheus, and a cooler site, Keppel (Figure 27A-B). In this way, 30 genotypes were simultaneously exposed to distinct, ecologically realistic conditions. We refer to corals replaced at their home sites as native samples (KK and OO samples), and corals placed at the alternative site as transplanted samples (KO and OK samples)(Figure 27A). Following a 3-

month acclimatization period, tissues were collected from each sample and assayed for gene expression using Tag-seq (Meyer et al. 2011), and DNA methylation using MBD-seq (Dixon et al. 2016). These data were analyzed in the context of fitness-related traits to assess the role of GBM in acclimatization. Specifically, we tested three predictions: 1) GBM changes in response to environmental conditions, 2) GBM covaries with fitness-related traits, and 3) changes in GBM covary with changes in gene expression. Although GBM overall remained remarkably consistent among fragments of the same genotype, we do find support for predictions 1 and 2, but not for 3. We conclude that GBM tracks important genome-environment interactions, but does not appear to directly regulate transcription.

METHODS

Reciprocal Transplantation Experiment

Field work was conducted with permission from the Great Barrier Reef marine Park Authority (Research permit G09/29894.1) as described previously (Dixon et al. 2014). Reciprocal transplantations were made between two environmentally distinct study sites (Keppel: 23°09S 150°54E and Orpheus 18°37S 146°29E) separated by 4.5 degrees of latitude in the Great Barrier Reef (Figure 27A). On the 23rd April (Orpheus) and 4th May (Keppels) 2010 fifteen colonies were collected from wild populations from each site and divided into two. One half of each colony was replaced in its native habitat, while the second half was transplanted to the alternate study site. Samples from all coral fragments were collected at midday after three months (9th July 2010 at Orpheus, 14th July at Keppels) frozen in liquid nitrogen, then transferred into RNAlater (Ambion, Austin, TX, USA) for gene expression and DNA methylation profiling.

MBD-seq library preparation

DNA was isolated from adult holobiont tissue using dispersion buffer (4M guanadine thiocyanate, 30mM sodium citrate, 30mM β-mercaptoethanol) followed by phenol chloroform purification and

a final cleanup with Zymo Genomic DNA Clean and Concentrator-10 kit (Catalog No D4011). Genomic DNA was sheared using a Misonix Sonicator 3000 to a size range of ~200 to 800 bp. Enrichment reactions were performed using the MethylCap kit (Diagenode Cat. No. C02020010) with an initial input of 2µg of sheared DNA per reaction. The methylated fraction was eluted from the capture beads in a single step using High Elution Buffer. Library preparation and sequencing of the enriched fragments was performed at the University of Texas Genome Sequencing and Analysis Facility. For a subset of 12 samples, both the enriched and the flow-through fractions were sequenced. Fold differences between these flowthrough libraries and their methylation enriched counterparts allowed us to assess absolute levels of methylation across genes.

MBD-seq data processing

Fifty MBD-seq libraries were prepared from reciprocally transplanted coral fragments. Sequencing produced 980 million raw reads with a mean of 16 ± 0.65 SEM million reads per sample. Raw reads were trimmed of non-template sequence using Cutadapt (Martin 2011) and quality filtered using Fastx toolkit (http://cancan.cshl.edu/labmembers/gordon/fastx_toolkit/). Adapter trimming and quality filtering reduced these the total read count 940 million, mean = 15 ± 0.64 SEM per sample. The reference genome and annotations (version 1.1) for *Acropora digitifera* (Shinzato et al. 2011) were downloaded from NCBI. To guard against reads originating from endosymbiont DNA contributing to gbM signal, the transcriptome sequences for *Symbiodinium* Clades A, B, C, and D (Seneca and Palumbi 2015) were appended to the end of the reference genome. Trimmed and filtered reads were mapped to this concatenated reference using Bowtie2 (Langmead and Salzberg 2012). To ensure that using a reference from an alternate species did not severely impair mapping, we compared the mean mapping efficiency against the *A. digitifera* reference with that of a draft genome sequence for *A. millepora* produced by David Miller and coworkers (James Cook University). Mean mapping efficiency against the *A. digitifera* (78.4 ± 0.7% SEM) reference was 5.2% lower than that of *A. millepora* (83.6 ± 0.8% SEM), hence genomic differences between the two species did not appear to substantially impair read alignment.

Following alignment, PCR duplicates were removed using Picard (https://broadinstitute.github.io/picard/). Mean duplication frequency was 14.3 ± 0.5% SEM. Mapped reads overlapping annotated coding sequences were counted using intersection-nonempty method in HTseq version 0.6.1p1 (Anders et al. 2015). As the genetic context of DNA methylation can be highly important to its effect on gene expression (Jones 2012), and methylation in the first exon can be inhibitory (Brenet et al. 2011), reads overlapping the window bounded from 1000bp upstream to 200bp downstream of the TSS were not counted as gene body methylation. We chose a 200bp downstream boundary based on the observation that reduction in MBD-scores (log_2 fold differences between captured and flowthrough samples) at the TSS extended roughly this far downstream into the gene body (Figure 28). This was done to ensure that signal due to methylation near the TSS was not attributed to gene body methylation. MBD-scores surrounding TSSs were calculated for 100bp windows. Fold coverage was counted using BEDTools (Quinlan and Hall 2010).

Tag-seq data processing

Transcription was quantified using Tag-seq (Meyer et al. 2011)(Lohman et al. 2016). It is important to note that while the DNA for MBD-seq and RNA for Tag-seq came from separate tissue samples from the same coral fragments. Hence correlations between methylation state and transcription reflect systemic phenotypes of the colony fragments. The tag-seq reads were downloaded from the SRA database (accession SRP049522; (Dixon et al. 2014)) and mapped against the *A.digitifera* genome using SHRiMP (Rumble et al. 2009). Mapped reads overlapping annotated coding sequences were counted using intersection-nonempty method in HTseq version 0.6.1p1 (Anders et al. 2015). Normalization or raw counts and statistical analyses were performed using DESEq2 (Love et al. 2014).

SNP calling

The MBD-seq reads were used to call SNPs for the 22 sequenced colonies. For this procedure, we concatenated reads from clone pairs into single files, and mapped the reads to a draft genome sequence for *A. millepora* (produced by David Miller and coworkers at James Cook University) using Bowtie2 (Langmead and Salzberg 2012). For the subset of 12 samples that we sequenced both captured and flowthrough fractions, both sets of reads were concatenated for SNP calling. Clone pairs were also called independently to serve as genotyping replicates for quality assessment. Genotyping was performed as described in (Dixon et al. 2015). One step of the GATK pipeline, variant quality score recalibration, requires a well-established set of SNPs in order to recalibrate quality scores to better fit these known SNPs. Lacking such a SNP set for *A. millepora*, we instead used a set of variants that with agreeing genotype calls across each set of replicates. After removal of singletons, the final set included 3713 variants. Based on comparison of genotype calls across replicates, accuracy (agreement across replicates) of this final set was 86.6%.

Assessing variation in gene body methylation and transcription

As with gene expression analyses, normalization and statistical analyses of MBD-seq reads were performed with DESEq2 (Love et al. 2014). As described in Dixon et al. (2016), we quantified absolute levels of GBM as the $\log_2$ fold difference in fold coverage between captured and flow-through libraries from a subset of 12 samples while controlling for genotype.

Tests for effects of origin and transplantation were performed in two ways. To assess overall associations, we used data from all samples and likelihood ratio tests comparing the full model (translation site + origin) to a reduced model that lacked the parameter of interest.

To more closely examine effects of origin and transplantation, additional tests were performed on samples subset by treatment. To test for origin effects, were performed two tests, comparing the two groups placed at Keppel to each other (OK vs KK) and comparing the two groups placed at Orpheus to each other (OO vs KO). These tests were intended to identify effects of origin while controlling for environmental conditions experienced during the experiment. To

assess effects of transplantation, we compared groups that originated from Orpheus to each other (OO vs OK) and compared groups that originated from Keppel to each other (KK vs KO). For these tests we included an additional parameter for colony identity (genotype). This comparison was intended to identify effects of transplantation while controlling for genotype. Results from these tests were used to identify genes showing evidence of plasticity that were input into downstream analyses.

Discriminant analysis of principal components

Plasticity of GBM patterns was further analyzed using discriminant analysis of principal components (DAPC) implemented in the R package adegenet (Jombart et al. 2010). DAPC is a multivariate analysis method designed to identify between-group variation while neglecting within-group variation. We used this method to distill highly multivariate genetic (SNPs), epigenetic (MBD-seq) and transcriptional (Tag-seq) datasets into single axes that maximized discrimination between natives from the two experimental sites (ie KK and OO samples). To focus on effects of plastic genes, only genes that showed evidence of transplant effects (raw $p < 0.01$; KO vs KK or OO vs OK comparisons) for were input into DAPC. This was true for both the MBD-seq and Tag-seq DAPC. After the discriminant function was fit to the native samples (OO and KK), it was applied to the transplanted samples (OK and KO). In this way, the loading values for the transplanted samples could be used to quantify how similar their patterns of transcription or methylation of plastic genes were to those typical of corals native to the transplantation site. We quantified this similarity for each transplanted sample as the inverse distance between the its loading value and the mean value for natives of the transplantation site (Fig. 4A). Specifically, we took the absolute value of the difference between each transplant's DAPC loading value and the mean value for native of the transplant site, converted these distances into z-scores, and multiplied the z-scores by -1 so that they reflect proximity to native patterns rather than distance. We refer to this value as the 'convergence', indicating the convergence toward the 'target' native patterns.

It seems reasonable that, either due to natural selection or plasticity, native corals would possess patterns optimal for their particular site. If this is true, then transplanted samples with patterns more similar to natives should show higher fitness. To test this prediction, we regressed the convergence values against five fitness-related traits: Daily weight gain, lipid content, carbohydrate content, protein content, and zooxanthellae density, as well as a summary fitness index (the first principal component for weight gain, lipid content, carbohydrate content, and protein content). The same analysis was performed for convergence values based on the discriminant functions from transcription and SNP data.

To further dissect the effect of GBM match on fitness proxies, we examined two separate aspects of convergence: pre-convergence, and shift. Pre-convergence was calculated similarly to convergence, only based on the distance between the non-transplanted clone member and mean for the alternative site. Pre-convergence thus quantified innate similarity in GBM patterns between each genotype and natives of the alternative site. Shift was calculated as the z-score for the difference in DAPC loading values for each transplanted fragment and its native clone-mate. This value therefore reflects the extent to which GBM patterns for each genotype diverged as a result of transplantation. We used AIC to show that models that included both pre-convergence and shift as predictors explained variation in fitness proxies better than convergence alone. To identify the optimal model for predicting fitness, we compared linear models with each combination of our genetic (SNPs), epigenetic (MBD-seq), and transcriptional (Tag-seq) variables along with interactions with origin. Based on AIC, the optimal model included GBM pre-convergence, GBM shift, and an interaction between GBM pre-convergence and origin. The relative importance of these factors was estimated using analysis of variance.

Validation of MBD-seq results with targeted bisulfite sequencing

To validate our MBD-seq results we used targeted bisulfite sequencing. Genomic DNA for this procedure was extracted and purified as before. It should be noted that these DNA extractions came from separate tissues from the MBD-seq and Tag-seq. We performed bisulfite conversion

using the EZ DNA Methylation-Gold kit (Zymo Research; cat. no. D5005). Thermocycler conditions for conversion were 98°C for 10 minutes followed by 53°C for 4 hours. Primers for post-conversion amplification were designed using Bisulfite Primer Seeker (Zymo Research; http://www.zymoresearch.com/tools/bisulfite-primer-seeker). We designed 13 primer sets to target coding sequences (excised from the *A. digitifera* reference genome) of 13 separate loci. Primer sequences including 5' tails for downstream addition of barcodes and Illumina adapters by PCR. Target loci were selected based on either strong origin or transplant effects in the DESEq2 analyses and or because they showed strong module membership in interesting modules from a Weighted Gene Coexpression Analysis (WGCNA; Langfelder and Horvath 2008). Amplification from bisulfite converted samples was performed using EpiMark Hot Start Taq DNA Polymerase (New England Biolabs; cat. no. M0490S). Thermocycler conditions were 95°C for 30 seconds followed by 35-40 cycles of 95°C for 15 seconds, 53°C for 30 seconds, 68°C for 30 seconds. Primer sequences are given in (Table 5). Barcodes and adapters for multiplexed sequencing were added to the resulting PCR products using PCR. The oligonucleotide sequences used for barcoding were the same as those given in the Tag-seq library preparation (Meyer et al. 2011; http://www.bio.utexas.edu/research/matz_lab/matzlab/Methods.html). Sequencing was performed using paired end 600 cycle runs on the Illumina Miseq. Resulting reads were quality trimmed using cutadapt (Martin 2011). Quantification of CpG methylation was performed using Bismark (Krueger and Andrews 2011) using a fasta file of the 13 exon sequences used to design the primers as a reference. CpG sites that were represented by fewer than 50 reads in a given sample were excluded from the analysis. To validate our estimates of absolute methylation levels, we calculated the mean percent methylation across all CpG sites within each gene across all samples, and regressed the log transformation of these values against the methylation scores from the MBD-seq data (Figure 27F). Six of the 13 genes selected showed effects of origin in the MBD-seq results. Of these, two were not sequenced for a sufficient number of samples from each site ($n<3$) for confident comparison between groups. For the remaining four, we reported differences in means between samples grouped by origin or all CpGs within the genes, and for each CpG individually.

We also showed that the difference in means based on origin for all CpGs in each gene correlated closely with the difference in mean normalized counts from the MBD-seq reads. The same analyses were performed for three genes showing evidence of transplantation effects in the MBD-seq results, with an additional assessment of pair-wise differences using clonal pairs. Differences due to transplantation measured with bisulfite sequencing were not significant, but were consistently in the same direction as indicated in the MBD-seq results.

Reporting of statistical results

Unless otherwise noted error bars for means reflect standard error of the mean. Adjustments for multiple test correction were performed using Benjamini Hochberg correction(Benjamini and Hochberg 1995). Adjusted p-values are reported using 'FDR' (eg FDR < 0.1). In many figures significance is indicated symbolically: ($^{n/s}$ not significant; $^{\&}$ $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$).

RESULTS

Absolute levels of GBM

For a subset of 12 samples, we sequenced both the captured and flow-through fractions from the MBD-seq library preparation. Log_2 fold differences between these samples were used to estimate absolute levels of methylation. As shown previously (Dixon et al. 2016), this measure was bimodally distributed across genes and correlated with normalized CpG content (CpGo/e)(Figure 27D-E). Targeted bisulfite sequencing of 13 loci further confirmed that MBD-seq accurately measures methylation in our system (Figure 27F).

GBM and transcription remains highly consistent among fragments of the same colony

Overall, patterns of GBM showed a strong dependence on colony identity. In spite of transplantation, all except one of the 22 clone-pairs showed greatest similarity to one another (Figure 29). Similar results were found for transcription (Figure 30), highlighting the importance

of genotype in shaping both methylation and gene expression patterns. Partitioning of variance between colony identity, origin, and transplantation site further confirmed these results, indicating an overwhelming effect of colony identity with only modest effects of origin and transplantation on both GBM and transcription (Figure 31C,G).

GBM linked with canalized transcription

Tests for differences in GBM depending on site of origin (irrespective of the site of transplantation) identified 197 differentially methylated genes (DMGs)(Figure 32A). Origin effects on methylation were validated using targeted bisulfite sequencing (Figure 33 and Figure 34). In terms of absolute methylation level, origin-specific DMGs tended to be intermediately or highly-methylated (Figure 35). Differential methylation by origin ($p < 0.01$) correlated positively with variation in transcription (Figure 32B): genes with higher GBM in one population tended to be more highly expressed in that population. This relationship was especially pronounced for genes that also tended toward differential transcription by origin ($p < 0.01$) (Figure 32C). Moreover, differential GBM between native fragments (OO vs KK) was correlated with transcription even among their transplanted clonal counterparts (OK vs KO) (Figure 36). This indicates that stable differences in GBM between populations are predictive of canalized expression differences.

GBM patterns predict fitness in novel environments

The effect of transplantation on GBM was subtle. Although many genes, (2167), showed significant differences in transcription (FDR < 0.1), only two genes passed false discovery correction for GBM (Figure 31A-B, E-F). Correlation between transplant effects based on MBD-seq and targeted bisulfite-seq was weaker than for origin effects (Figure 34), but differences were generally in the same direction (Figure 37). In terms of absolute methylation, genes that tended toward site-specific methylation (raw P <0.01) tended to be weakly methylated (Figure 35). Seventeen of these genes also showed a tendency toward origin-specific methylation.

To better examine these subtle environmental effects, we used discriminate analysis of principal components (DAPC). DAPC is designed to find the axis in multivariate space that best discriminates samples into predefined groups (Jombart et al. 2010). The function that describes this axis can then be applied to values from additional samples to assess their variation in the context of the pre-specified contrast (Kenkel and Matz 2016). We used DAPC to discriminate between native samples (KK and OO; Figure 27A) based on genes that showed evidence of GBM plasticity ($p < 0.01$ transplant effect; see methods). We then applied the discriminant function to the foreign transplants (Figure 31D). The same analysis was performed using transcriptional data (Figure 31H), and for SNP data (Figure 38). Based on both the number of significant genes (FDR < 0.1), and the magnitude of shift along with discriminant axis (Figure 31), transcription was much more plastic than GBM.

Projection of our transplanted samples onto the discriminant axis allowed us to quantify the extent to which the transplants' GBM patterns matched those of native corals. Initially, we found that daily weight gain correlated with DAPC coordinates, but only of transplanted samples (Figure 39B). The nearly orthogonal relationships for the two transplant groups suggested that greater 'convergence' toward native GBM patterns predicted greater fitness. To further investigate this trend, we calculated a convergence value for each transplant that was inversely proportional to the distance along the discriminant axis between the transplant and the mean for natives of the site (Figure 39A; see methods). We then regressed these convergence values against fitness-related traits. Strikingly, five different traits (percent daily weight gain, lipid concentration, carbohydrate concentration, protein concentration, and density of zooxanthellae) correlated positively with convergence (Figure 40C-G). The same analyses were performed using transcription data (Figure 41), and SNP data (Figure 38), but did not detect any significant relationships. To provide a summary index for coral fitness, we took the first principal component (explaining 44% of variation) for four of the fitness proxies (weight gain, lipids, carbohydrates, and protein) among the transplanted samples (Figure 42). This fitness index also correlated with convergence (Figure 39C).

To further dissect the nature of these relationships, we examined two contributing components of convergence: *pre-convergence* and *shift*. Pre-convergence was similar to convergence, but calculated based on the non-transplanted clone mate rather than the transplant itself (Figure 39A). Pre-convergence is intended to describe how similar a colony's GBM were to the local mean prior to transplantation. Shift was calculated as the proportional distance along the discriminant axis between each transplanted sample and its native clone mate. Shift was intended to describe the extent of plastic change in GBM in response to transplantation. Based on AIC, the linear model that included both pre-convergence and shift provided better prediction of fitness than convergence alone (AIC = 0.75 and 1.87 respectively). Comparing a diversity of linear models, including predictors from the SNP and transcription discriminant axes, we found that the optimal linear model for fitness included pre-convergence, shift, and an interaction between pre-convergence and origin (AIC = -2.05). Of these predictors, pre-convergence explained greatest amount of fitness variation (Figure 39D).

DISCUSSION

GBM is a signature for canalized transcription

Within plant and animal genomes, correlations between GBM and transcription are generally weak (Zhang et al. 2006; Ball et al. 2009; Wang et al. 2014; Dixon et al. 2016), and evidence that GBM directly regulates transcription in a general context remains scarce (Zilberman 2017). Some associations however, are consistent. Across plant and animal taxa, GBM is often bimodally distributed, separating genes into strongly and weakly methylated classes (Takuno and Gaut 2012; Sarda et al. 2012). In both cases, strongly methylated genes tend toward moderately elevated transcription across broad cellular, developmental, and ecological contexts, whereas weakly methylated genes tend toward context specificity. Here we show that in a basal metazoan, variation in GBM between populations is predictive of variation in transcription. Genes with elevated GBM in one population tend to show higher transcription in that population, even when

the individuals are transplanted to alternative environments. These results further establish GBM as a signature for stably active transcription, and demonstrate that variation in GBM between populations could potentially be of functional importance.

GBM and acclimatization

Changes in GBM in response to transplantation were subtle, considerably weaker than the gene expression response, and continued to be predominantly attributable to genetics (broad-sense heritability). However, analysis of genes showing trends (raw p=0.01) for GBM change upon transplantation showed that for transplanted corals, convergence of GBM patterns toward those of native corals positively correlated with fitness. Convergence of GBM patterns had two components, the extent to which the colony already matched its target (pre-convergence), and the extent to which its GBM patterns changed during the experiment (shift). Our results indicate that GBM shift explains roughly 20% of variation in fitness, with roughly 50% explained by pre-convergence and pre-convergence by site interactions. Because of the high similarity in GBM patterns between clone mates (Figure 29), we suggest that pre-convergence largely reflects genetic diversity, and that shift is the better measure of GBM plasticity. Indeed, pre-convergence significantly correlated with its equivalent measure based on SNP data (Figure 43). With this in mind, roughly 20% of fitness variation was explained by GBM plasticity, with an additional 50% likely explained by genetic pre-adaptation (Figure 39D). Previous work on another species of *Acropora* emphasized the importance of acclimatization for thermal adaptation (Palumbi et al. 2014). These findings suggest that comparative methylation assays can shed light on the extent to which corals are acclimatized to particular environments.

Missing mechanism

While our results demonstrate a clear association between GBM and fitness, the actual mechanism linking DNA methylation with phenotype remains unclear. The third prediction of our hypothesis was that environmentally induced changes in GBM would covary with transcription.

Although population-specific GBM and transcription were correlated (Figure 32), GBM patterns associated with transplantation showed either no correlation, or a *negative* correlation with transcription (Figure 44 and Figure 45). Hence, in this dataset, the relationship between environmentally dynamic GBM and transcription was qualitatively different from that observed for origin-based differences and across genes within plant and animal genomes (Zilberman et al. 2007; Zemach et al. 2010) genomes. One explanation for this inconsistency concerns the fact that the RNA for our transcription assays and the DNA for our methylation assays came from separate tissue samples. Any covariation between transcription and GBM therefore depended on the consistency of both processes across the coral colony. If origin-based differences in transcription tend to be more consistent across colonies than environmentally responsive genes, it could explain the differing relationships with GBM. Another alternative is that the dynamics of GBM and transcription operate on distinct timescales. We suggest that GBM changes slowly, only in response to sustained changes in transcription. If this is the case, GBM patterns could provide a more integrated picture expression across long time periods, in contrast to the temporally localized 'snapshot' provided by the transcriptome. This could potentially explain the surprising result that GBM convergence predicted fitness (Figure 39; Figure 40) when transcriptional convergence did not (Figure 41), and why GBM correlates with stable, origin-based transcription but not with environmental dynamic transcription. To clarify, the transcriptional differences we observed likely included not only responses to conditions *characteristic* of the two sites, but also to transient conditions, such as the weather immediately preceding collection. Such transient conditions could produce differences in transcription disproportionate to their actual ecological importance. If, on the other hand, accumulation or depletion of GBM results from persistent changes in transcription, it may reflect genome-environment interactions with greater importance for fitness. This hypothesis could be tested with time series of concurrent GBM and transcriptome assays, or experimental paradigms known to cause particular persistent changes in transcription. In any case, more work is needed to unravel the causal relationships between GBM and transcription.

Conclusions and outlook

Here we present four major results. First, patterns of GBM and transcription depend predominantly on genotype. This result highlights the need to carefully consider genotypic effects in interpretations of ecological transcriptomic and methylomic data. Second, differences in GBM between populations correlated with similar differences in transcription, demonstrating that variation GBM not only predicts transcriptional activity within genomes, but also between populations. Third, GBM is considerably less plastic than transcription. As a result, large sample populations are likely needed to detect significant environmental effects on GBM. Finally, patterns of GBM correlate with coral fitness under ecologically realistic novel conditions. This result demonstrates the potential for methylomics to provide insight into highly complex ecological interactions and inform management strategy.

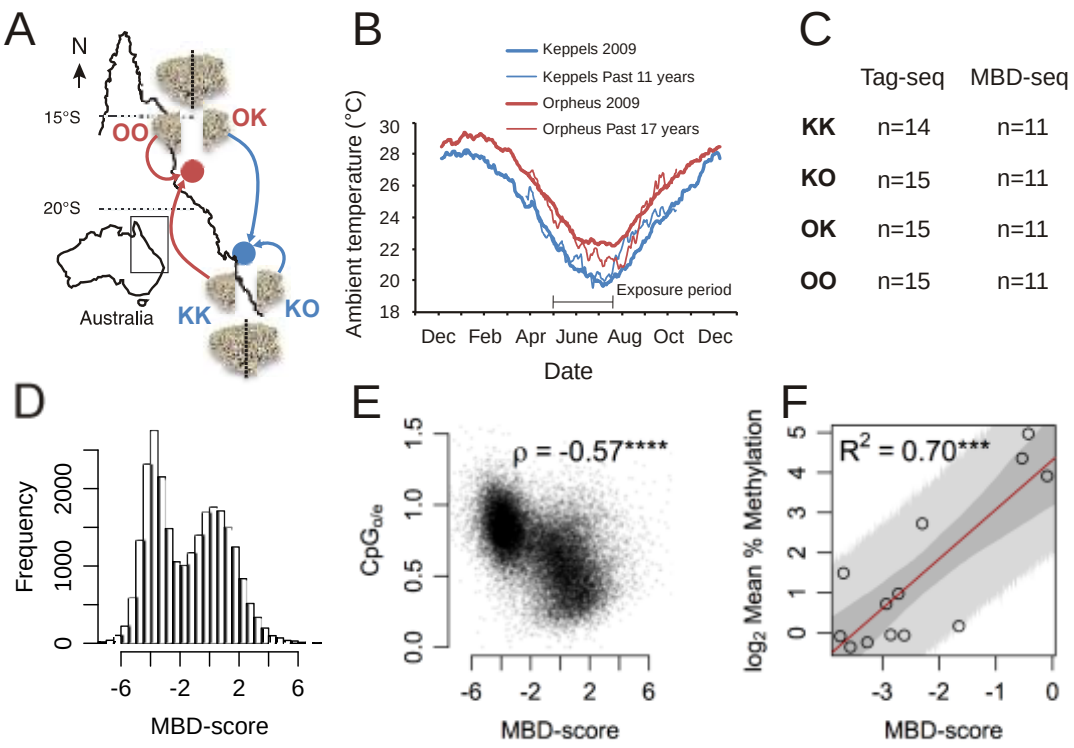

Figure 27 Experimental design and validation of MBD-seq. (A) Map of experiment location in the Great Barrier Reef, Australia. Colonies were divided into fragments and reciprocally transplanted between two sites, a northern site Orpheus (red), and a southern site Keppel (blue). Sample groups are labeled with first letter indicating origin and second letter indicating transplant location (eg KO samples originated from Keppel and were transplanted to Orpheus). (B) Ambient temperatures differ between the two sites, providing distinct environmental pressures. (C) Table of sample sizes for transcription (Tag-seq) and methylation (MBD-seq) assays. (D) Distribution of methylation level (MBD-score) for all genes. MBD-score was calculated as the $\log_2$ fold difference between paired captured and flow-through libraries from the MBD-seq library preparation (n=12 pairs; see methods). Bimodal distribution of these values is consistent with expectations for GBM in invertebrate species. (E) Correlation between methylation score and normalized CpG content (CpGo/e), a metric that reflects historical germline methylation known to correlate with somatic methylation in diverse invertebrates (Sarda et al. 2012). (F) Correlation between methylation estimates based on MBD-seq and targeted bisulfite sequencing. Mean percent methylation was calculated as the proportion methylated CpG sites within each gene averaged across all samples. Red line traces the expectation for linear model. Grey shading indicates 90% posterior probability intervals for the mean (darker), and sample distribution (lighter).

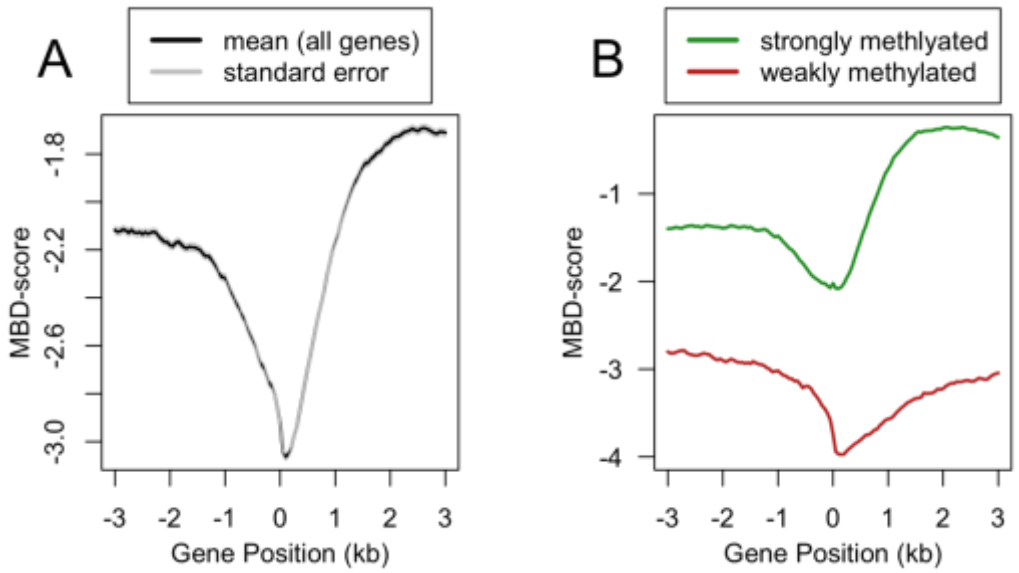

Figure 28 MBD-seq fold coverage reduced at transcription the start sites (TSS) and into gene bodies. Regions 3 Kb upstream and downstream from transcription start sites were divided into 100 bp windows (30 windows upstream and downstream with one central window spanning the TSS). Fold coverage within these windows was counted using bedtools.

Table 5 Primer sequences used for targeted bisulfite sequencing. Each primer includes a 5' tail (bold) for amplification with barcoded primers for multiplex Illumina sequencing using in the Tag-seq library preparation. Sequences for these oligoes are available here: https://github.com/z0on/tag-based_RNAseq.

Primer	Sequence (5'-3')
LOC107327073_bsfTag_For	**CTACACGACGCTCTTCCGATCT**TTTTTGYGAGGTTGATTTTGTTATTATG
LOC107327073_bsfTag_Rev	**ACGTGTGCTCTTCCGAT**TTTCCAAACATATTCCTTTCCATAACATTC
LOC107356898_bsfTag_For	**CTACACGACGCTCTTCCGATCT**TAGATTTYGTTATAATGTTATTAAGAAGTGAAGG
LOC107356898_bsfTag_Rev	**ACGTGTGCTCTTCCGAT**AATAATATTAAACATTCTCTCTACAAATCTACCAC
LOC107358158_bsfTag_For	**CTACACGACGCTCTTCCGATCT**TGGTYGGATTGTTGAAGAGTTTAAGTAG
LOC107358158_bsfTag_Rev	**ACGTGTGCTCTTCCGAT**ACACCCAAATCACCCATCTCATTAAC
LOC107356899_bsfTag_For	**CTACACGACGCTCTTCCGATCT**GTTTTTAAAATTTGAAGGATTTGGTTTTGTTG
LOC107356899_bsfTag_Rev	**ACGTGTGCTCTTCCGAT**TTACATTATATTTTCCAAACATATTTCATACCATAAC
LOC107358871_bsfTag_For	**CTACACGACGCTCTTCCGATCT**GATATYGGGTTTTTAATAATAATTGTATGTTGGTTG
LOC107358871_bsfTag_Rev	**ACGTGTGCTCTTCCGAT**TTTTCCTTTAAAATTATTTCCACCCAACTCC
LOC107334334_bsfTag_For	**CTACACGACGCTCTTCCGATCT**ATAAGGATATGTAGGGTTTTGGTAAGG
LOC107334334_bsfTag_Rev	**ACGTGTGCTCTTCCGAT**AATTCATAATAACTACCCTACAACAAAAAATCCC
LOC107347512_bsTag_For	**CTACACGACGCTCTTCCGATCT**ATYGATAGATTAAAAGAAGTTGGAGTATTG
LOC107347512_bsTag_Rev	**ACGTGTGCTCTTCCGAT**CCTAAACAATCCATAAAAACCTTCCTACAATTC
LOC107350794_bsTag_For	**CTACACGACGCTCTTCCGATCT**GAYGTGTTTTGTATTTAGTTATTGGATATTTGG
LOC107350794_bsTag_Rev	**ACGTGTGCTCTTCCGAT**ATCTTCCRCAATTAAAAAAACATACTTAAAATCTTCAC
LOC107339795_bsTag_For	**CTACACGACGCTCTTCCGATCT**TGAATATAGTTAAGGTAAATGGATGAGTTATATATGG
LOC107339795_bsTag_Rev	**ACGTGTGCTCTTCCGAT**CAAAAATAAACRAATCAAACAAAACAATCATTAC
LOC107336909_bsTag_For	**CTACACGACGCTCTTCCGATCT**AGAGATGYGGAATAGATATTTTTGGTTGG
LOC107336909_bsTag_Rev	**ACGTGTGCTCTTCCGAT**AATAACRAACATTACATCTTATTTCTCTAAAATAATAC
LOC107352877_bsTag_For	**CTACACGACGCTCTTCCGATCT**TGGTGTYGTGAATTTGTTTAAATAATTTATG
LOC107352877_bsTag_Rev	**ACGTGTGCTCTTCCGAT**CATTCCATCRACACAACTAAATAAAAAAC
LOC107351808_bsTag_For	**CTACACGACGCTCTTCCGATCT**GTYGAGATATGTTAAAATTTGAGGATGTGAGTTTG
LOC107351808_bsTag_Rev	**ACGTGTGCTCTTCCGAT**ATCCACCATTCRCCAACTTACTAATAAATCTCCC
LOC107327285_bsTag_For	**CTACACGACGCTCTTCCGATCT**ATTATGTTAGYGTTTTTGTTATTTTGGATTGTGG
LOC107327285_bsTag_Rev	**ACGTGTGCTCTTCCGAT**CCTCATTCACAAAAACAACCTTTAATC

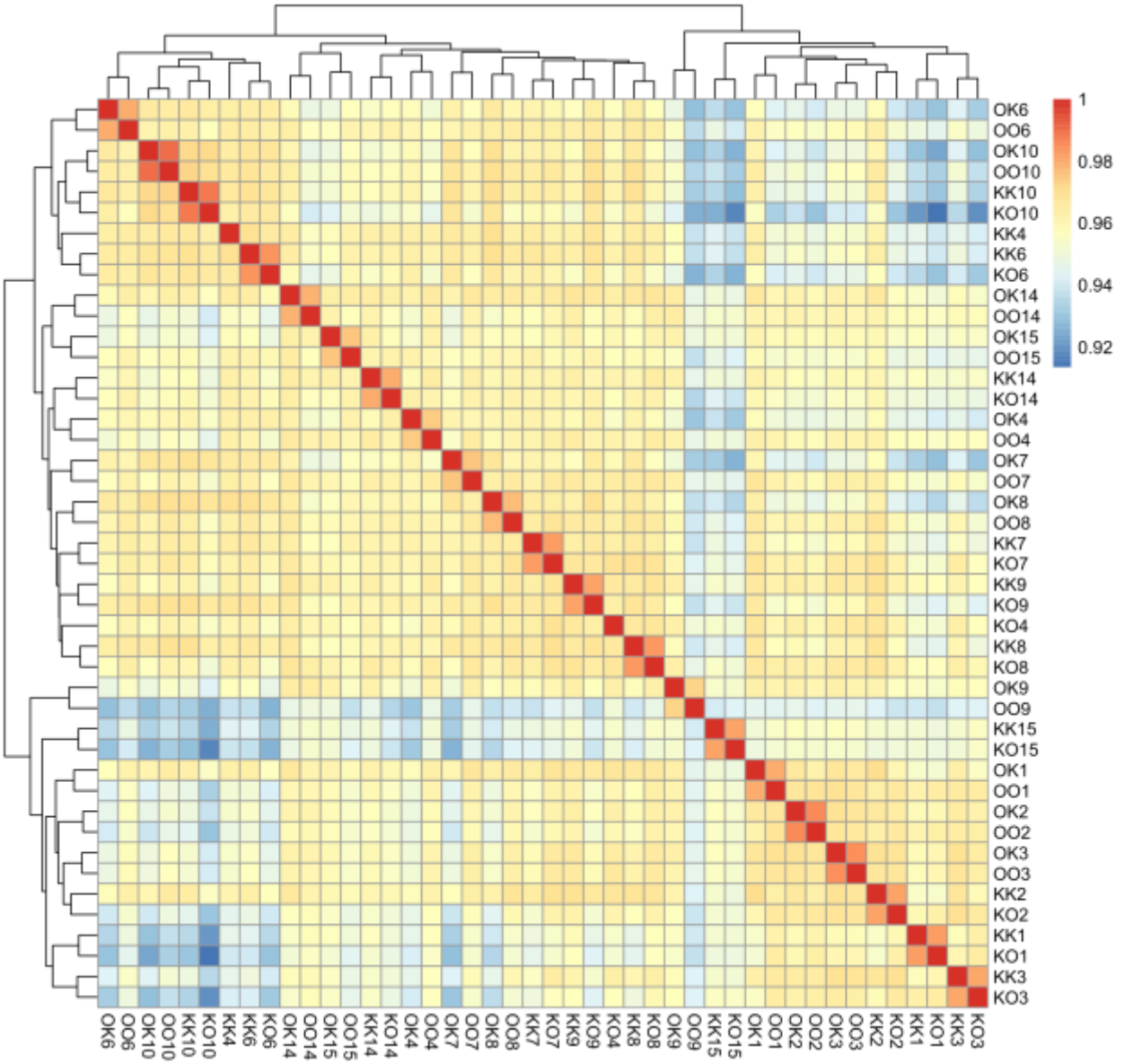

Figure 29 Heatmap of overall correlations of gene body methylation patterns illustrating strong genetic component. Colors indicate Spearman's rank correlations for normalized MBD-seq read counts across all coding genes (N=24001). Samples were clustered by maximum distance. First letter of sample names indicates sample origin. Second letter indicates transplantation site. Number indicates replicate. Samples sharing the same first letter and the same number are clonal fragments from the same colony. All of the 22 clone pairs except one were most similar to one another.

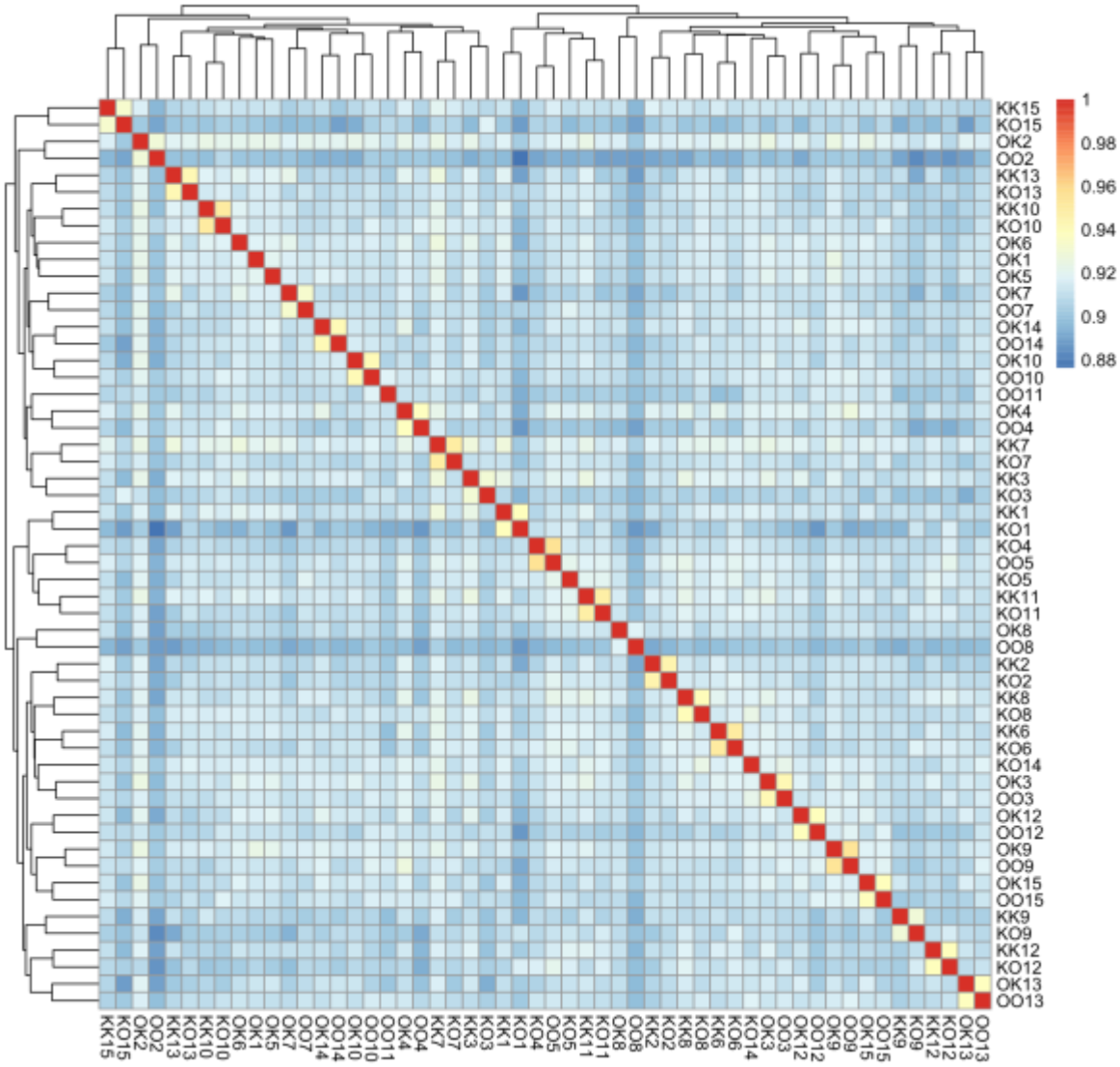

Figure 30 Heatmap of overall correlations for transcription illustrating strong genetic component. Colors indicate Spearman's correlations for normalized Tag-seq read counts across all coding genes (N=19706). Samples were clustered by maximum distance. Samples were clustered by maximum similarity. First letter of sample names indicates sample origin, second letter indicates transplantation site, number indicates replicate. Samples sharing the same first letter and number are clone fragments from the same colony. All of the 24 clone pairs except were most similar to one another (six samples lacked data for clone pairs).

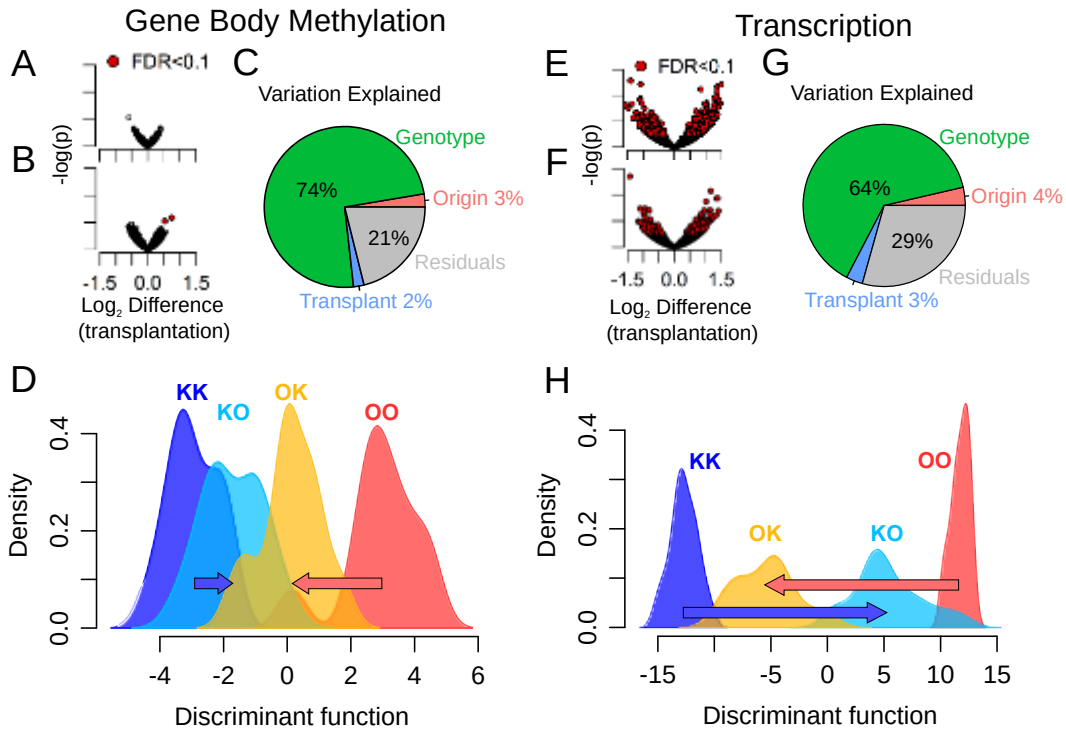

Figure 31 Effects of transplantation on gene body methylation and transcription. (A) Summary of effects transplantation on GBM for all genes (n= of corals from Keppel (KK vs KO). (B) Effect of transplantation on GBM of corals from Orpheus (OO vs OK). (C) Density plot of sample loading values for discriminant analysis of principal components (DAPC). Normalized read counts for genes showing evidence of GBM plasticity ($p < 0.01$ in either of transplantation tests) were input into DAPC to discriminate between the native groups (KK and OO). The function was then applied to read counts from the transplanted groups (KO and OK). Loading values for the transplanted fragments summarize the shift in their GBM patterns to more resemble those of native corals. Arrows indicate the change in mean loading values from each native group to their transplanted clonal counterparts. (D-F) The same figures generated based on transcription (Tag-seq).

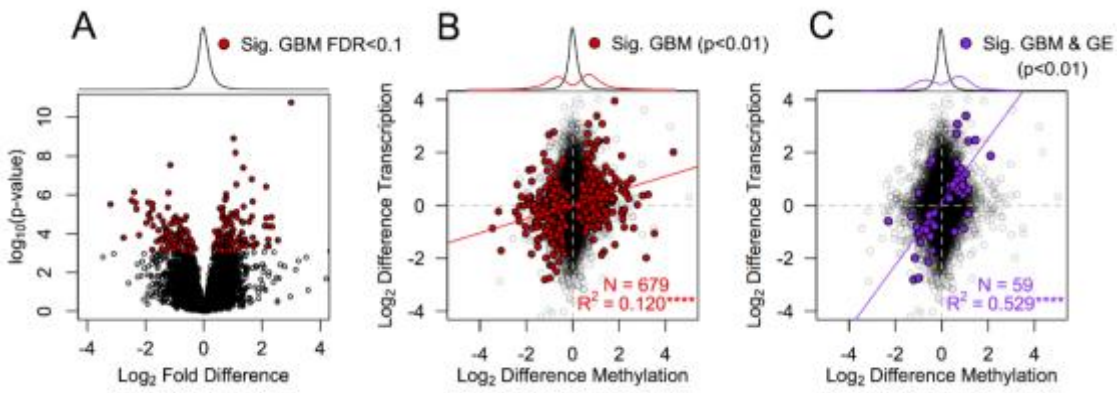

Figure 32 Correlation between origin-specific GBM and transcription. (A) Differential GBM between all fragments from Keppel and all fragments from Orpheus. Significant genes (FDR < 0.1) are shown in red. (B) Scatterplot of log_2 fold differences in transcription and GBM. Log_2 fold differences are based on all fragments from Orpheus and all fragments from Keppel (OO and OK vs KK and KO). All genes are shown in black. Genes showing tendency ($p<0.01$) for origin-based differences in GBM are shown in red. The red line traces least squares regression for only these genes. (C) The same scatterplot illustrating the correlation of log_2 fold differences for genes showing tendency ($p<0.01$) for origin-based differences in both GBM and transcription (purple). Purple line traces least squares regression for these genes. Traces above each scatterplot indicate x-axis density for all points (black) or overlaid points as indicated by color. Asterisks indicate significance of linear regressions (**** $p < 0.0001$).

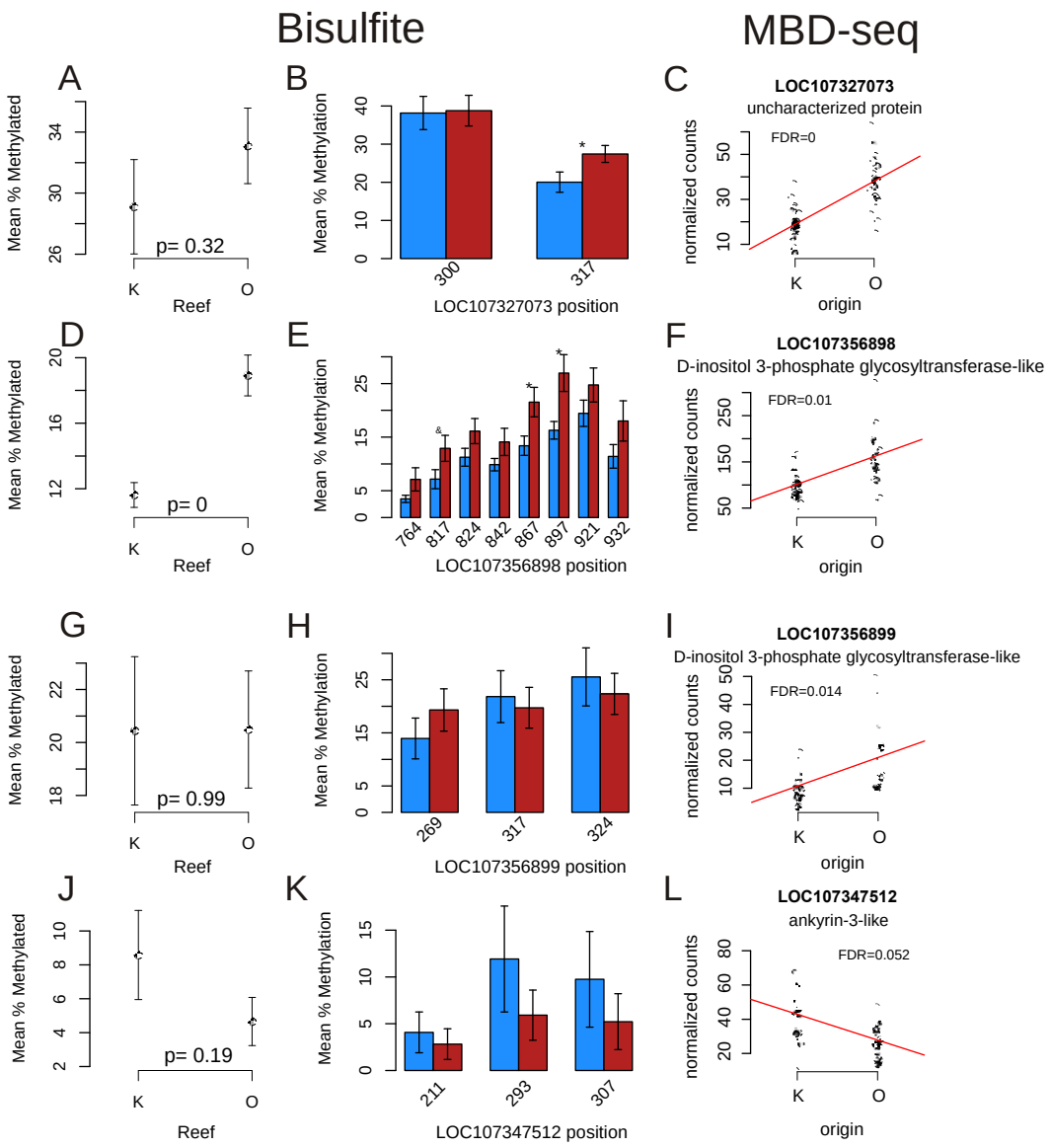

Figure 33 Comparison of origin effects detected with targeted bisulfite sequencing and MBD-seq. Each row shows a separate locus. Column 1 shows mean percent methylation across all CpG sites within the locus for site of origin. Column 2 shows the mean percent methylation of each CpG site individually by site of origin. Column 3 shows normalized read counts from the MBD-seq results. Error bars indicate standard error of the mean. P-values indicate significance based on Student's t-tests (* $p < 0.05$; & $p < 0.1$). Six loci showing origin effects in the MBD-seq results were assayed for DNA methylation with targeted bisulfite sequencing. Two of these were not sequenced for a sufficient number of samples from each site ($n<3$) for confident comparison. Of the remaining four, three showed differences in the same direction as indicated in the MBD-seq results, and two demonstrated significant differences in at least one CpG site.

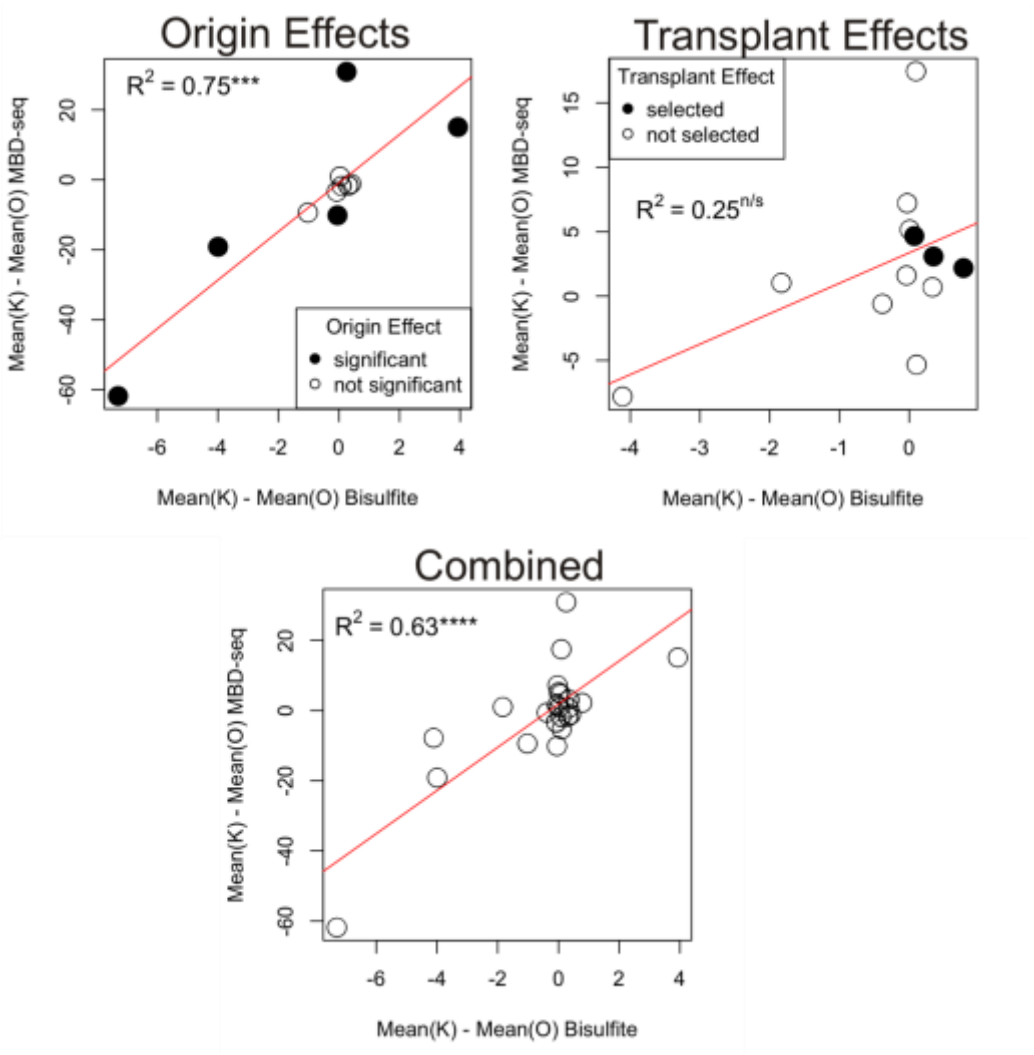

Figure 34 Validation of origin effects using targeted bisulfite sequencing. Thirteen selected loci were assayed for DNA methylation using targeted bisulfite sequencing. Plots show regressions of mean normalized read count against mean percent methylation across all CpG sites for each locus (see methods) split either by origin, transplantation site, or both. Red lines indicate least squared regressions (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$).

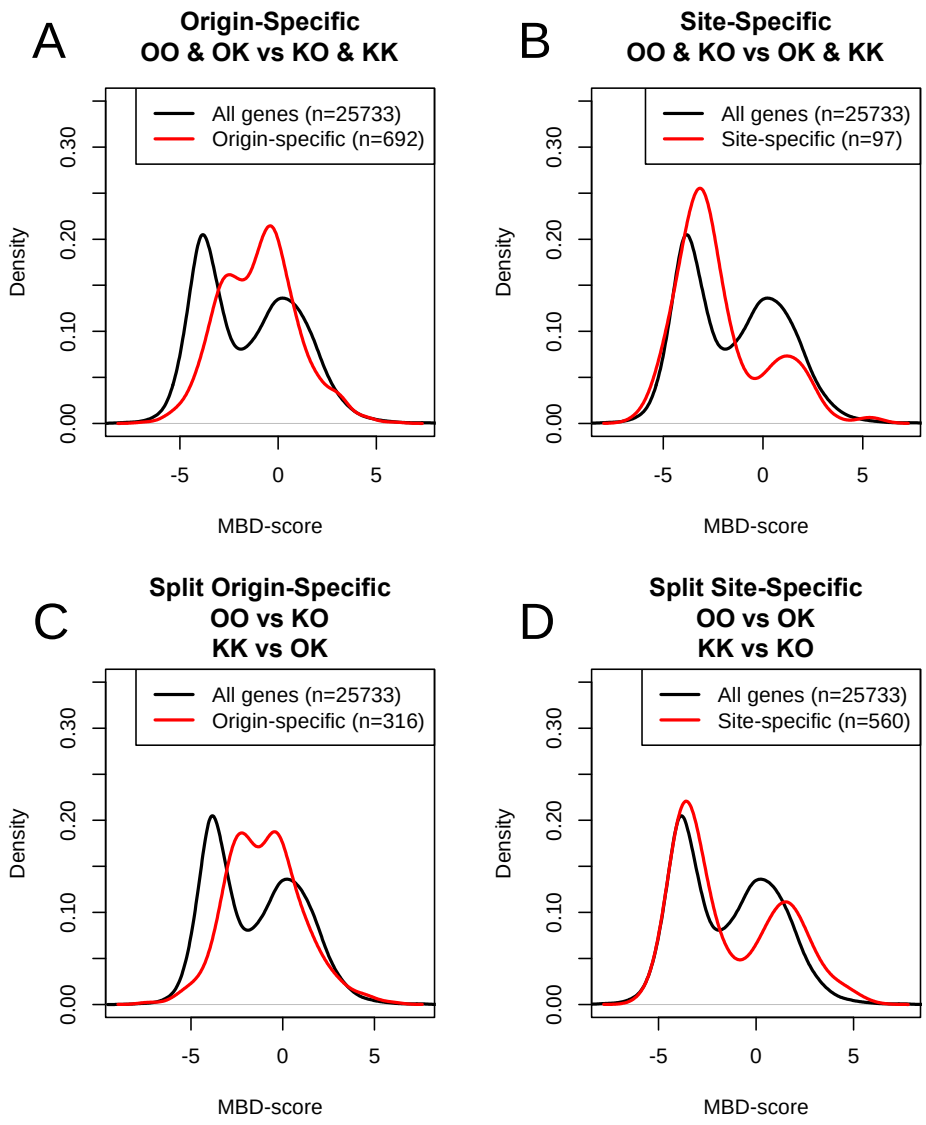

Figure 35 Distribution of absolute methylation levels (MBD-score; see methods) for genes showing origin and site-specificity based on the full dataset and on split models. For each plot, the density of MBD-scores for all genes is shown in black and the density for the indicated subset is shown in red. (A) Density for genes showing tendency toward origin-specific methylation for full dataset model (raw $p < 0.01$; testing for differences between all samples originating from Orpheus and all samples originating from Keppel). (B) Density for genes showing a tendency toward site-specific methylation for full dataset model (raw $p < 0.01$; testing for differences between all samples placed at Orpheus and all samples placed at Keppel). (C) Density for genes showing tendency toward origin specific methylation in split dataset models (raw $p < 0.01$; testing for differences based on origin among samples placed at the same site during the experiment). (D) Density for genes showing tendency toward site-specific methylation in split dataset models (raw $p < 0.01$; testing for differences based on transplantation within populations).

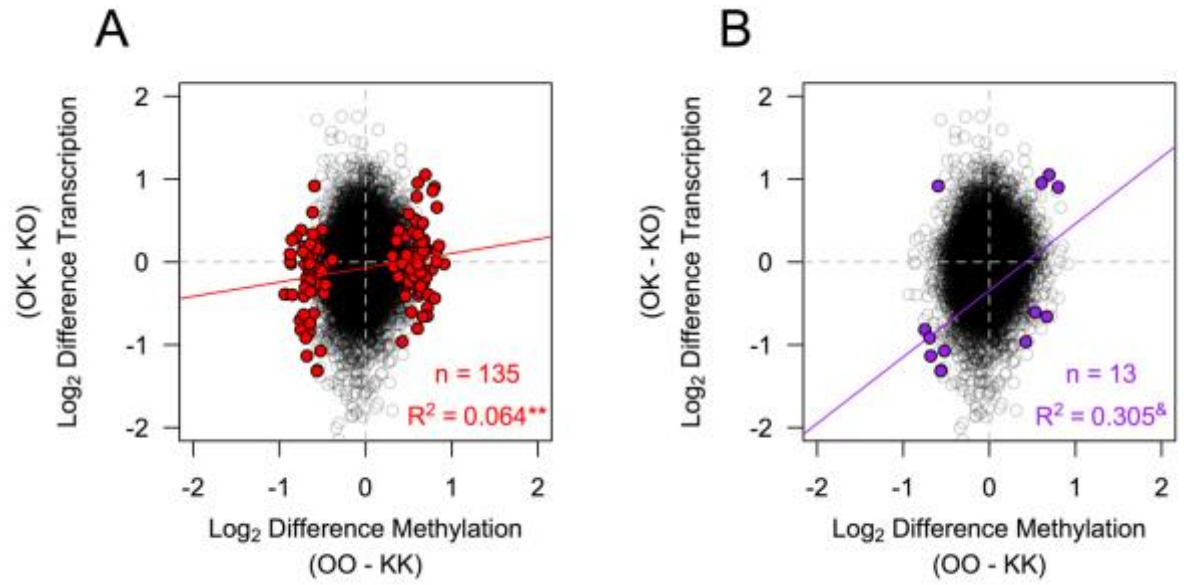

Figure 36 Differential GBM between native corals correlates with origin-based differences among transplanted counterparts. Scatterplots show log_2 fold differences in GBM between native samples (x-axis) and log_2 fold differences in transcription for transplanted samples (y-axis). In both plots data points for all genes are shown in black (A) Genes that showed a tendency ($p < 0.01$) toward differences in GBM between native samples are shown in red. Red line traces least squares regression for these genes. (B) Genes that showed a tendency ($p < 0.01$) toward differences in GBM and transcription are shown in purple. Purple line traces least squares regression for these genes. Symbols indicate significance (& $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$).

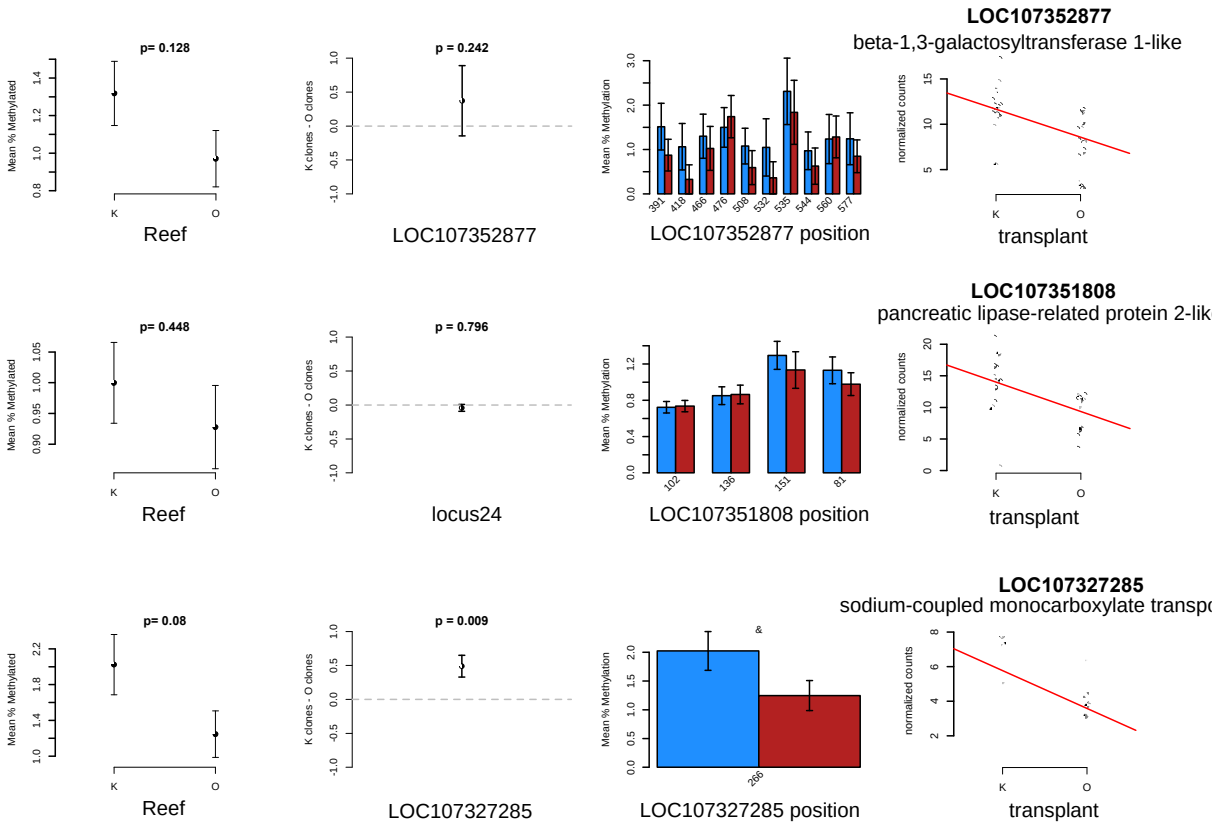

Figure 37 Comparison of transplant effects detected with targeted bisulfite sequencing and MBD-seq. Column 1 shows mean percent methylation across all CpG sites within the locus for transplantation site. Column 2 shows the mean difference between fragments placed at Keppel (KK and OK samples) and their clone pairs placed at Orpheus (OO and KO samples). Column 3 shows the mean percent methylation of each CpG site individually by transplantation site. Column 4 shows normalized read counts from the MBD-seq results. Error bars indicate standard error of the mean. P-values indicate significance based on Student's t-tests (* $p < 0.05$; & $p < 0.1$). All three loci show variation in bisulfite results in the same direction as indicated in the MBD-seq results.

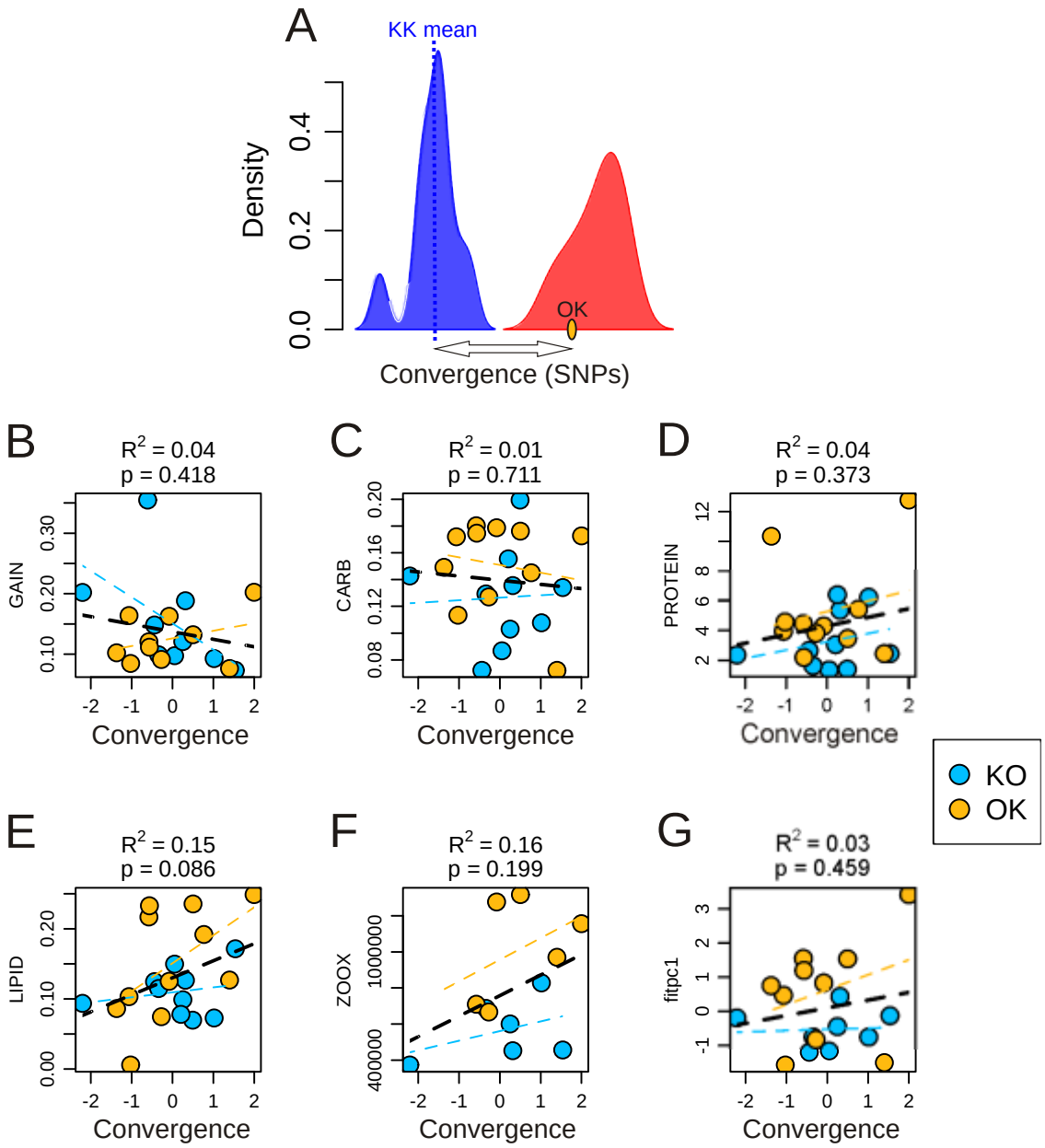

Figure 38 Correlation between SNP convergence score and physiology. Discriminant analysis of principal components was conducted based on SNP data generated from the MBD-seq reads (see methods). SNPs were called for each colony (rather than fragment pairs) so only two distributions are shown. Match score was calculated as the inverse distance of a given colony from the mean value for corals native to the site. We detected no significant relationships between transcription match score and physiological measures. Thick dotted lines trace least squares regression for all data points. R^2 and p-values above each panel refer to this linear model. Thin dotted lines trace least squares regressions for each transplantation group individually.

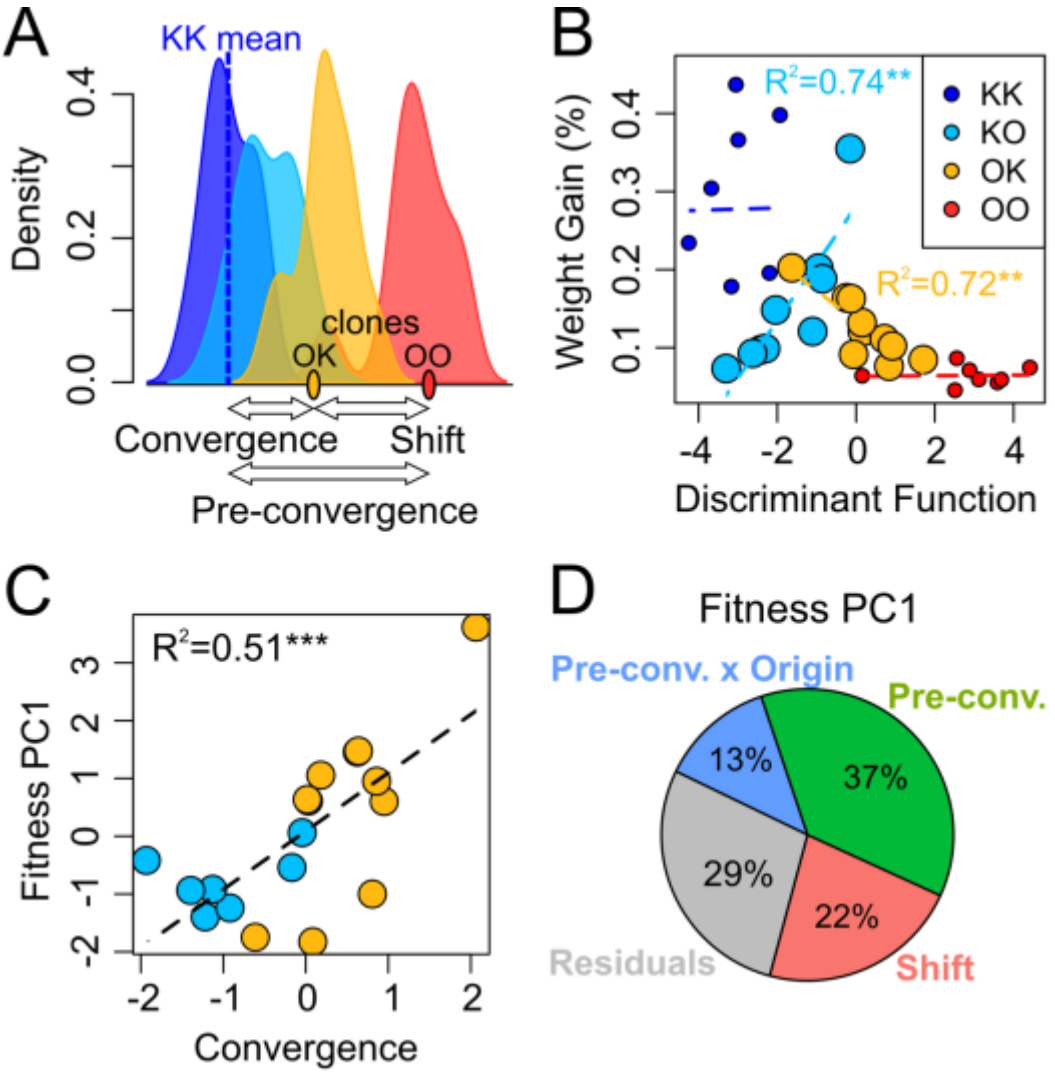

Figure 39 Correlation between gene body methylation (GBM) and physiology. (A) Projection of transplanted samples onto the discriminant axis allowed us to quantify the degree to which GBM patterns in transplants converged on those typical of native corals. Convergence was calculated as the inverse distance of a transplant from the mean value for corals native to the site. Match score could be described as two separate components: 'shift' which describes how much the transplanted sample's GBM patterns shifted from its native clonal counterpart, and pre-convergence, which describes how similar the genotype already was to the native mean for transplantation site (see methods). (B) Scatterplot showing correlation between samples' discriminant axis coordinates and daily percent weight gain. The nearly orthogonal relationships seen for the two transplant groups shows how convergence of their GBM patterns toward those of native corals was associated with higher growth rates, an important fitness proxy for stony corals. (C) Correlation between convergence and a summary fitness index: the first principal component (44% of variance explained) for gain, lipid, carbohydrate, and protein. (D) Pie chart showing analysis of variance results for optimal linear model of PC1 from (C).

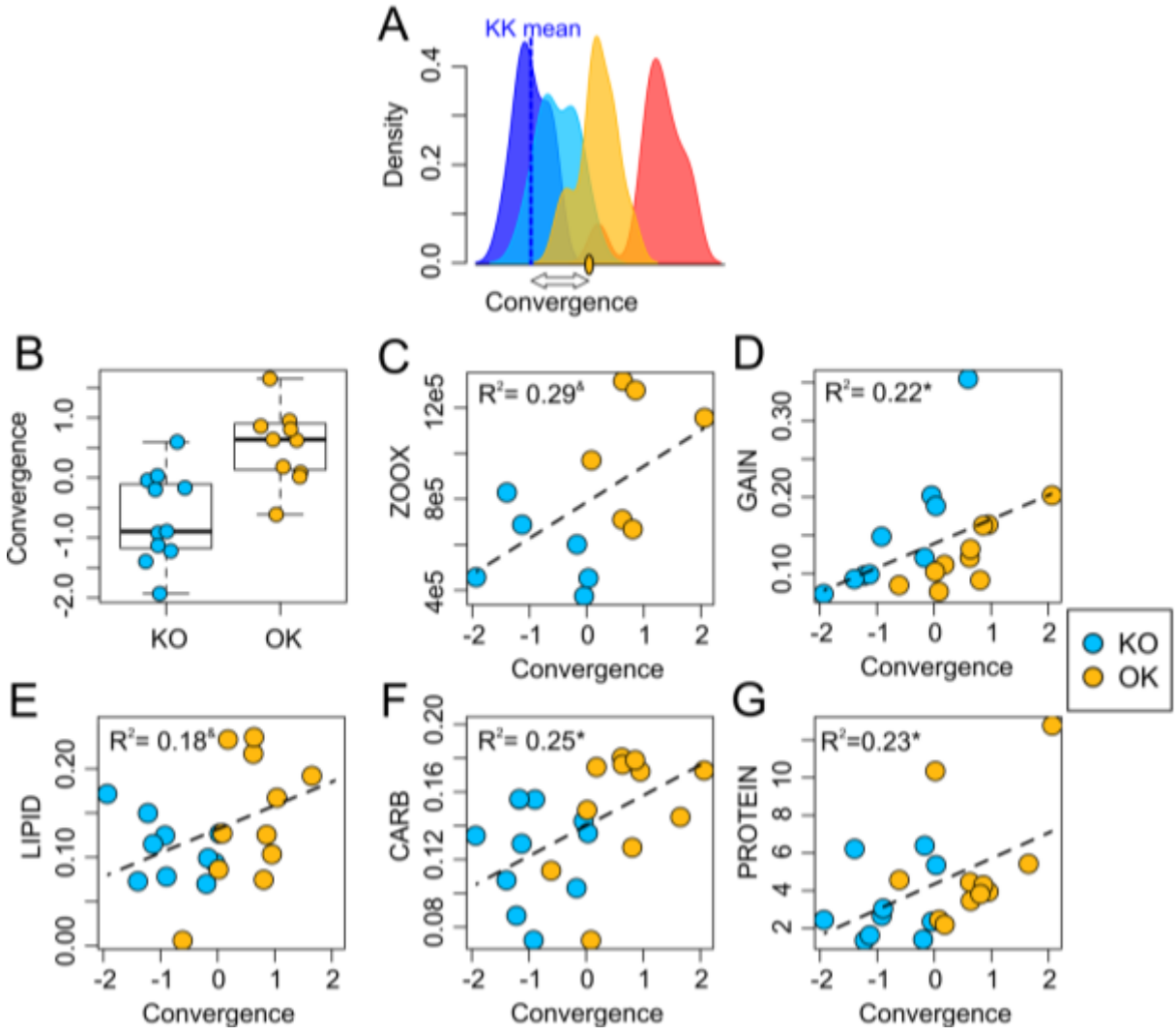

Figure 40 Correlation between gene body methylation (GBM) and physiology. (A) Projection of transplanted samples onto the discriminant axis allowed us to quantify the degree to which GBM patterns in transplants converged on those typical of native corals. Convergence was calculated as the inverse distance of a transplant from the mean value for corals native to the site. Match score could be described as two separate components: 'shift' which describes how much the transplanted sample's GBM patterns shifted from its native clonal counterpart, and pre-convergence, which describes how similar the genotype already was to the native mean for transplantation site (see methods). (B) Scatterplot showing correlation between samples' discriminant axis coordinates and daily percent weight gain. The nearly orthogonal relationships seen for the two transplant groups shows how convergence of their GBM patterns toward those of native corals was associated with higher growth rates, an important fitness proxy for stony corals. (C) Correlation between convergence and a summary fitness index: the first principal component (44% of variance explained) for gain, lipid, carbohydrate, and protein. (D) Pie chart showing analysis of variance results for optimal linear model of PC1 from (C).

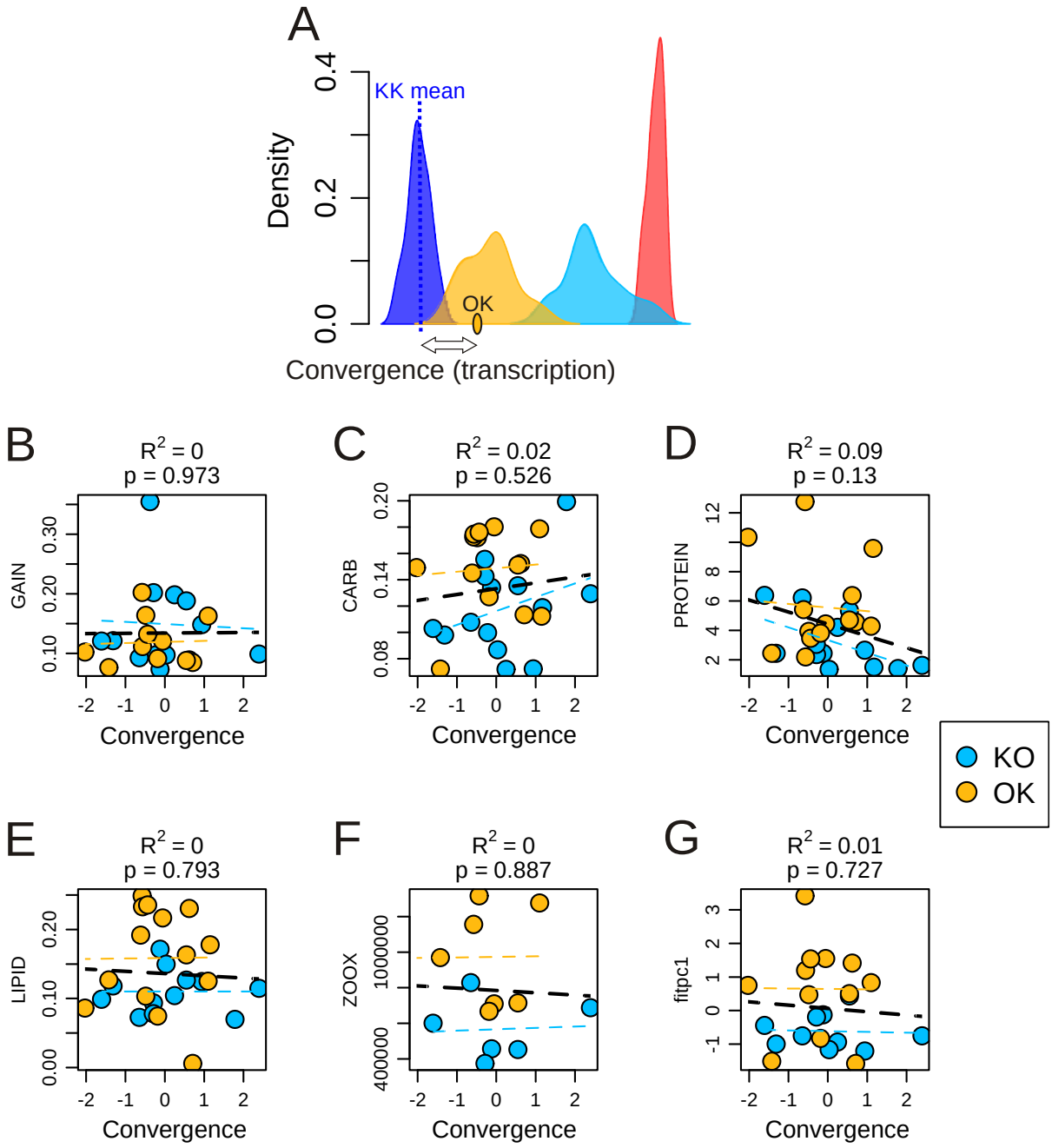

Figure 41 Correlation between transcription convergence score and physiology. Projection of transplanted samples onto the discriminant axis allowed us to quantify the degree to which GBM patterns in transplants matched those typical of native corals. Match score was calculated as the inverse distance of a transplant from the mean value for corals native to the site. We detected no significant relationships between transcription match score and physiological measures. Thick dotted lines trace least squares regression for all data points. R^2 and p-values above each panel refer to this linear model. Thin dotted lines trace least squares regressions for each transplantation group individually.

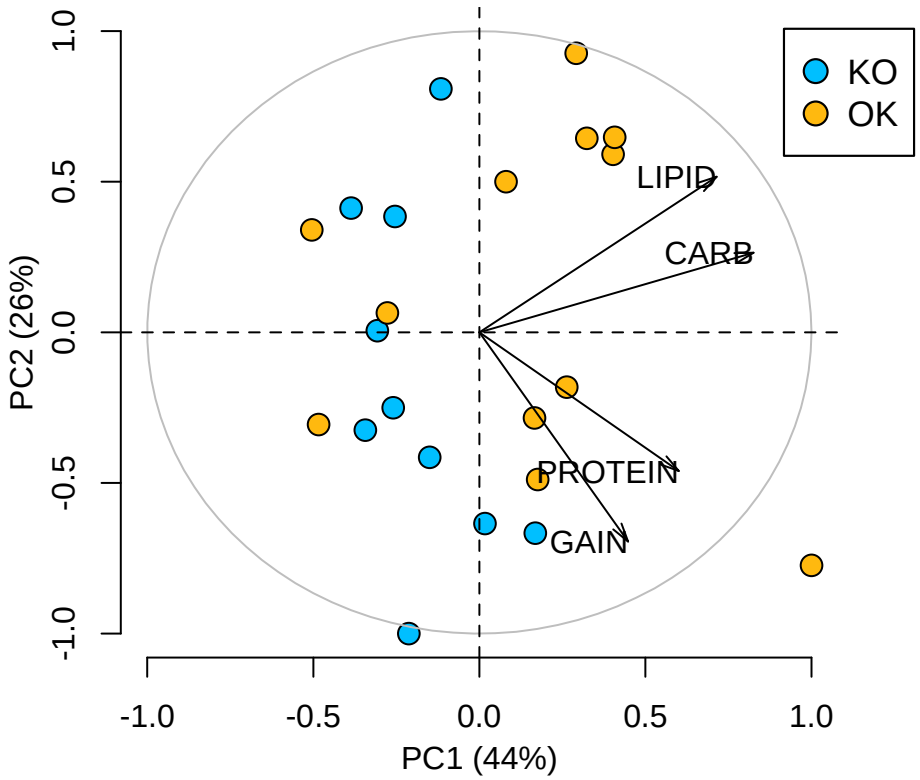

Figure 42 Principal component analysis of four physiological measures used as fitness proxies. The first component (explaining 44% of variation) was used as a summary index for fitness.

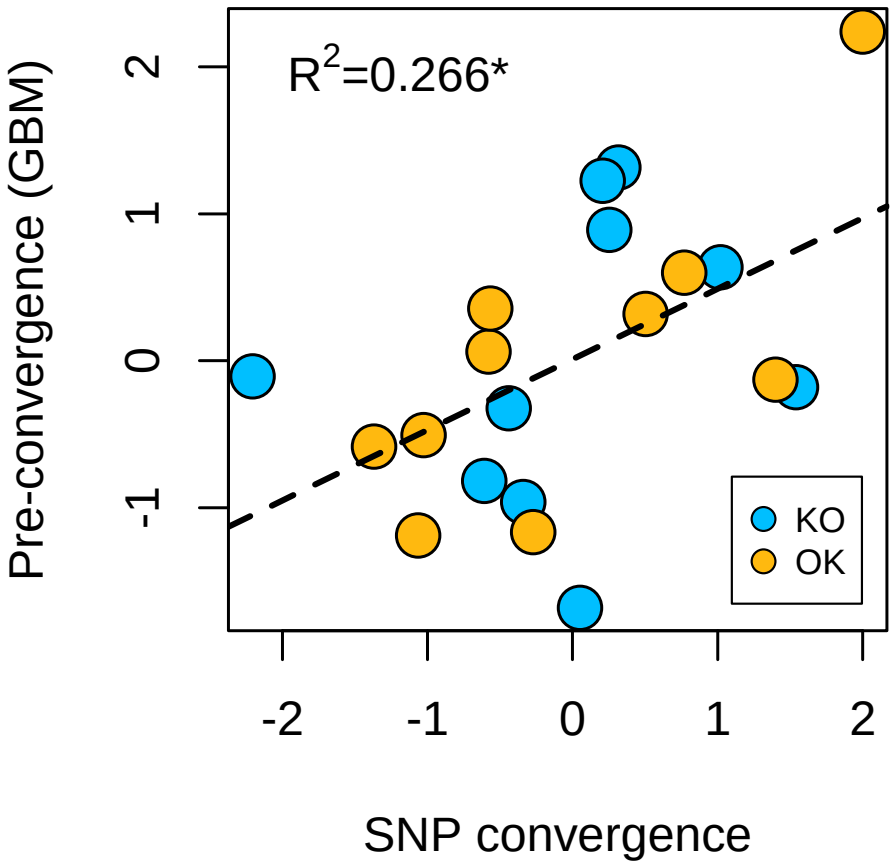

Figure 43 Correlation between pre-convergence and SNP convergence. Black dotted line traces least squares regression. Asterisk indicates significance (* $p < 0.05$). Positive correlation here indicates that pre-convergence is driven by genotype.

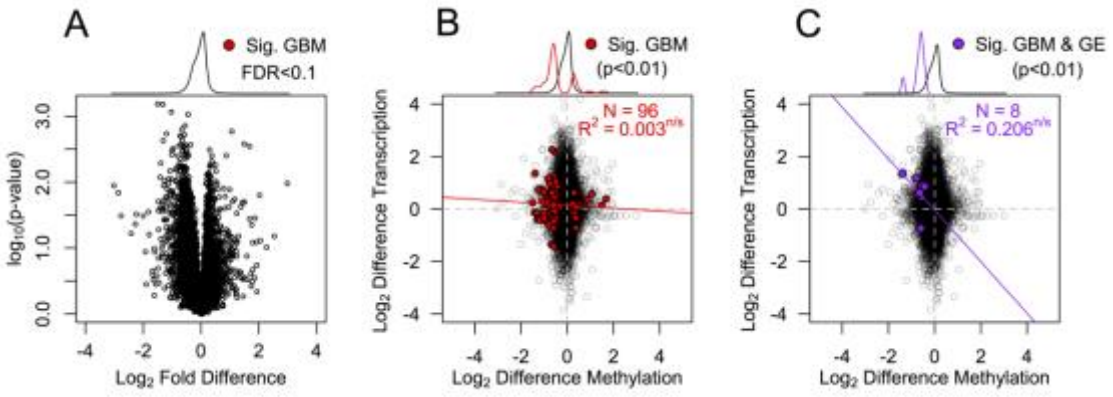

Figure 44 No correlation detected between transplantation site-specific GBM and transcription. (A) Differential GBM between all fragments placed at Keppel and all fragments placed at Orpheus during the experiment. No genes passed false discovery correction (FDR<0.1). (B) Scatterplot of log_2 fold differences for transcription and GBM. Log_2 fold differences are based on all fragments placed at Orpheus and all fragments placed at Keppel (OO and KO vs KK and OK). All genes are shown in black. Genes that showed a tendency ($p < 0.01$) toward site-specific differences in GBM are shown in red. The red line traces least squares regression for only these genes. (C) The same scatterplot illustrating log_2 fold differences for genes that showed a tendency ($p < 0.01$) toward site-specific differences for both GBM and transcription (shown in purple). Purple line traces least squares regression for these genes. Traces above scatterplots indicate density for all datapoints and the overlaid subset indicated by color. These traces highlight the fact that methylation tended to be higher among samples placed at Keppel (negative log_2 fold differences).

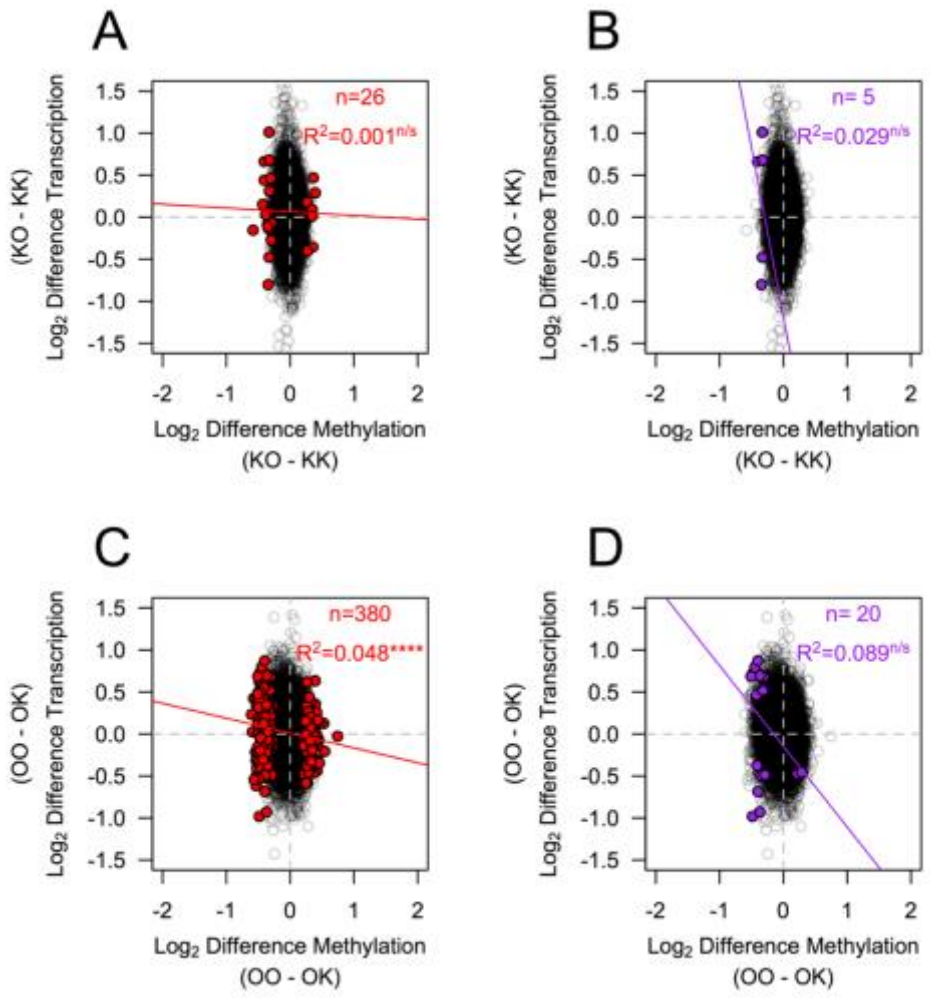

Figure 45 Correlation between environmentally induced changes in GBM and transcription. Scatterplots show $\log_2$ differences due to transplantation site in transcription (Y-axis) and GBM (X-axis). All data points are shown in black. (A,C) Genes that showed a tendency ($p < 0.01$) toward differences in GBM due to transplantation are shown in red. Red line traces least squares regression for these genes. (B,D) Genes that showed a tendency ($p < 0.01$) toward differences due to transplantation in GBM and transcription are shown in purple. Purple line traces least squares regression for these genes. Symbols indicates significance (& $p < 0.1$; * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$).

CONCLUSION

This dissertation can be summarized with three major conclusions: 1) populations in the Great Barrier Reef have the potential for rapid genetic adaptation to climate change 2) patterns of DNA methylation predict gene expression plasticity 3) patterns of gene body methylation can predict coral fitness under environmental change. Of these, the third is the most novel and surprising. The adaptive function of GBM remains surprisingly and stubbornly uncertain. Indeed, many question whether GBM has any adaptive function at all (Bewick and Schmitz 2017). I, and others have argued that this is unlikely, as GBM comes at a high mutational cost, and in genomic regions where they are most likely to be deleterious (Dixon et al. 2016; Zilberman 2017). Surely, if GBM served no function, lineages that lost the machinery required to establish and maintain it would become more common, and supersede those that continue to bear its useless metabolic and mutational costs. It could be argued that this is simplistic, and that GBM is a pleiotropic effect of useful methylation elsewhere, such as promoter methylation or inhibition of transposable elements. Evidence from Hymenopteran species contradicts this however, as they show no evidence of regulatory promoter methylation (Zemach et al. 2010), and are actually *hypo*methylated at transposable elements (Libbrecht et al. 2016). Moreover, the associations between GBM and gene expression patterns described in Chapter 2 show clear similarities even with plants and other animals. This is either a remarkable instance of convergent evolution, or indicates that GBM been conserved in Eukaryotic lineages for over a billion years.

But what is the function? Prevailing theories include 1) regulation of splicing and 2) inhibition of spurious transcription initiation. Several studies report evidence for the former (Flores et al. 2012; Libbrecht et al. 2016), and hence regulation of splicing currently has the greatest support. I have suggested that GBM accumulates in response to persistently active transcription.

Following the logic described in chapter 2, that actively transcribed genes pose a greater danger for deleterious errors during expression, it is reasonable that tighter regulation of splicing and inhibition of spurious initiation would both be of special adaptive value in these genes. Regrettably, the datasets presented here do not help to distinguish between these hypotheses. They do however, shed light on the interrelations between GBM and transcription, and its potential functional importance. Notably, in chapter 3 I show that variation in GBM between populations predicts variation in transcription. Moreover, after transplantation, resemblance to GBM patterns of native corals predicts higher fitness. These two findings are consistent with functional relevance of GBM, and at a minimum indicate a link between GBM and functionally relevant genomic processes.

REFERENCES

Aceituno, F. F., N. Moseyko, S. Y. Rhee, and R. a Gutiérrez. 2008. The rules of gene expression in plants: organ identity and gene body methylation are key factors for regulation of gene expression in Arabidopsis thaliana. BMC Genomics 9:438.

Akashi, H. 1994. Synonymous codon usage in Drosophila melanogaster: Natural selection and translational accuracy. Genetics 136:927–935.

Altschul, S. F., T. L. Madden, a a Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–402.

Altschul, S., W. Gish, and W. Miller. 1990. Basic Local Alignment Search Tool. J Mol Biol. 215:403–410.

Anders, S., and W. Huber. 2010. Differential expression analysis for sequence count data. Genome Biol. 11:R106. BioMed Central Ltd.

Anders, S., P. T. Pyl, and W. Huber. 2015. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169.

Angers, B., E. Castonguay, and R. Massicotte. 2010. Environmentally induced phenotypes and DNA methylation: how to deal with unpredictable conditions until the next generation and after. Mol. Ecol. 19:1283–95.

Aubin-Horth, N., and S. Renn. 2009. Genomic reaction norms: using integrative biology to understand molecular mechanisms of phenotypic plasticity. Mol. Ecol. 18:3763–80.

Ayre, D. J., and T. P. Hughes. 2000. Genotypic diversity and gene flow in brooding and spawning corals along the Great Barrier Reef, Australia. Evolution (N. Y). 54:1590–1605.

Ball, M. P., J. B. Li, Y. Gao, J.-H. Lee, E. M. LeProust, I.-H. Park, B. Xie, G. Q. Daley, and G. M. Church. 2009. Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat. Biotechnol. 27:361–8.

Barshis, D. J., J. T. Ladner, T. A. Oliver, F. O. Seneca, N. Traylor-Knowles, and S. R. Palumbi. 2013. Genomic basis for coral resilience to climate change. Proc. Natl. Acad. Sci. U. S. A. 110:1387–92.

Batut, P., A. Dobin, C. Plessy, P. Carninci, and T. R. Gingeras. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. 23:169–80.

Baubec, T., D. F. Colombo, C. Wirbelauer, J. Schmidt, L. Burger, A. R. Krebs, A. Akalin, and D. Schübeler. 2015. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature 520:243–7.

Behura, S. K., and D. W. Severson. 2013. Codon usage bias: Causative factors, quantification methods and genome-wide patterns: With emphasis on insect genomes. Biol. Rev. 88:49–61.

Benaglia, T., D. Chauveau, D. R. Hunter, and D. S. Young. 2007. Mixtools : An R Package for Analyzing Finite Mixture Models.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. 57:289–300.

Bewick, A. J., and R. J. Schmitz. 2017. Gene body DNA methylation in plants. Curr. Opin. Plant Biol. 36:103–110. Elsevier Ltd.

Bonasio, R., Q. Li, J. Lian, N. Mutti, L. Jin, H. Zhao, P. Zhang, P. Wen, H. Xiang, Y. Ding, Z. Jin, S. S. Shen, Z. Wang, W. Wang, J. Wang, S. L. Berger, J. Liebig, G. Zhang, and D.

Reinberg. 2012. Genome-wide and Caste-Specific DNA Methylomes of the Ants Camponotus floridanus and Harpegnathos saltator. Curr. Biol. 22:1755–1764.
Brenet, F., M. Moh, P. Funk, E. Feierstein, A. J. Viale, N. D. Socci, and J. M. Scandura. 2011. DNA methylation of the first exon is tightly linked to transcriptional silencing. PLoS One 6.
Bulmer, M. 1987. Coevolution of codon usage and transfer RNA abundance. Nature 325:278–730.
Capuano, F., M. Mülleder, R. Kok, H. J. Blom, and M. Ralser. 2014. Cytosine DNA methylation is found in Drosophila melanogaster but absent in Saccharomyces cerevisiae, Schizosaccharomyces pombe, and other yeast species. Anal. Chem. 86:3697–702.
Chapman, J., E. Kirkness, O. Simakov, S. E. Hampson, T. Mitros, W. Thomas, T. Rattei, P. G. Balasubramanian, J. Borman, D. Busam, K. Disbennett, C. Pfannkoch, N. Sumin, G. G. Sutton, L. D. Viswanathan, B. Walenz, D. M. Goodstein, U. Hellsten, T. Kawashima, S. E. Prochnik, N. H. Putnam, S. Shu, B. Blumberg, C. E. Dana, L. Gee, D. F. Kibler, L. Law, D. Lindgens, D. E. Martines, J. Peng, P. A. Wigge, B. Bertulat, C. Guder, Y. Nakamura, S. Ozbek, H. Watanabe, K. Khalturin, G. Hemmrich, A. Franke, R. Augustin, S. Fraune, E. Hayakawa, S. Hayakawa, M. Hirose, J. S. Hwang, K. Ikeo, C. Nishimiya-Fujisawa, A. Ogura, T. Takahashi, P. R. H. Steinmetz, X. Zhang, R. Aufschnaiter, M.-K. Eder, A.-K. Gorny, W. Salvenmoser, A. M. Heimberg, B. M. Wheeler, K. J. Peterson, A. Böttger, P. Tischler, A. Wolf, T. Gojobori, K. A. Remington, R. L. Strausberg, J. C. Venter, U. Technau, B. Hobmayer, T. C. G. Bosch, T. W. Holstein, T. Fujisawa, H. R. Bode, C. N. David, D. S. Rokhsar, and R. E. Steele. 2010. The dynamic genome of Hydra. Nature 464:592–596.
Chevin, L., S. Collins, and F. Lefèvre. 2013. Phenotypic plasticity and evolutionary demographic responses to climate change : taking theory out to the field. Funct. Ecol. 27:966–979.
Coleman-Derr, D., and D. Zilberman. 2012. Deposition of Histone Variant H2A.Z within Gene Bodies Regulates Responsive Genes. PLoS Genet. 8.
Consortium, T. U. 2013. Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res. 41:43–47.
Danecek, P., A. Auton, G. Abecasis, C. a Albers, E. Banks, M. a DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, and R. Durbin. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–8.
Davies, S. W., A. Marchetti, J. B. Ries, and K. D. Castillo. 2016. Thermal and pCO2 stress elicit divergent transcriptomic responses in a resilient coral. Front. Mar. Sci. 3:Article 122.
Davies, S. W., M. Rahman, E. Meyer, E. A. Green, E. Buschiazzo, M. Medina, and M. V Matz. 2013. Novel polymorphic microsatellite markers for population genetics of the endangered Caribbean star coral, Montastraea faveolata. Mar. Biodivers. 43:167–172.
De'ath, G., K. E. Fabricius, H. Sweatman, and M. Puotinen. 2012. The 27-year decline of coral cover on the Great Barrier Reef and its causes. Proc. Natl. Acad. Sci. 109:17995–17999.
DePristo, M. a, E. Banks, R. Poplin, K. V Garimella, J. R. Maguire, C. Hartl, A. a Philippakis, G. del Angel, M. a Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–8.
Dimond, J. L., and S. B. Roberts. 2016. Germline DNA methylation in reef corals: patterns and potential roles in response to environmental change. Mol. Ecol. 25:1895–1904.
Dixon, G. B., L. K. Bay, and M. V Matz. 2014. Bimodal signatures of germline methylation are

linked with gene expression plasticity in the coral Acropora millepora. BMC Genomics 15:1109.
Dixon, G. B., S. W. Davies, G. V Aglyamova, E. Meyer, L. K. Bay, and M. V Matz. 2015. Genomic determinants of coral heat tolerance across latitudes. Science (80-.). 348:1460–1462.
Dixon, G., L. K. Bay, and M. V Matz. 2016. Evolutionary consequences of DNA methylation in a basal metazoan. Mol. Biol. Evol. 33:2285–2293.
Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U. S. A. 102:14338–43.
Drummond, D. A., and C. O. Wilke. 2008. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell 134:341–352.
Duan, X., S. Yi, X. Guo, and W. Wang. 2015. A comprehensive analysis of codon usage patterns in blunt snout bream (Megalobrama amblycephala) based on RNA-Seq data. Int. J. Mol. Sci. 16:11996–12013.
Dunlap, W. C., A. Starcevic, D. Baranasic, J. Diminic, J. Zucko, R. Gacesa, M. J. van Oppen, D. Hranueli, J. Cullum, and P. F. Long. 2013. KEGG orthology-based annotation of the predicted proteome of Acropora digitifera: ZoophyteBase - an open access and searchable database of a coral genome. BMC Genomics 14:509. BMC Genomics.
Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Opin. Genet. Dev 12:640–649.
Duret, L., and D. Mouchiroud. 2000. Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol. Biol. Evol. 17:68–74.
Elango, N., B. G. Hunt, M. A. D. Goodisman, and S. V Yi. 2009. DNA methylation is widespread and associated with differential gene expression in castes of the honeybee, Apis mellifera. Proc. Natl. Acad. Sci. U. S. A. 106:11206–11211.
Falckenhayn, C., B. Boerjan, G. Raddatz, M. Frohme, L. Schoofs, and F. Lyko. 2013. Characterization of genome methylation patterns in the desert locust Schistocerca gregaria. J. Exp. Biol. 216:1423–1429.
Feil, R., and M. F. Fraga. 2011. Epigenetics and the environment: emerging patterns and implications. Nat. Rev. Genet. 13:97–109. Nature Publishing Group.
Fisher, R. A. 1922. On the Interpretation of $\chi 2$ from Contingency Tables , and the Calculation of P. J. R. Stat. Soc. 85:87–94.
Fisher, R. A. 1932. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
Flores, K. B., and G. V Amdam. 2011. Deciphering a methylome : what can we read into patterns of DNA methylation? J. Exp. Biol. 214:3155–3163.
Flores, K., F. Wolschin, J. J. Corneveaux, A. N. Allen, M. J. Huentelman, and G. V Amdam. 2012. Genome-wide association between DNA methylation and alternative splicing in an invertebrate. BMC Genomics 13:1. BMC Genomics.
Foden, W. B., S. H. M. Butchart, S. N. Stuart, J.-C. Vié, H. R. Akçakaya, A. Angulo, L. M. DeVantier, A. Gutsche, E. Turak, L. Cao, S. D. Donner, V. Katariya, R. Bernard, R. a Holland, A. F. Hughes, S. E. O'Hanlon, S. T. Garnett, C. H. Sekercioğlu, and G. M. Mace. 2013. Identifying the world's most climate change vulnerable species: a systematic trait-based assessment of all birds, amphibians and corals. PLoS One 8:e65427.
Foret, S., R. Kucharski, M. Pellegrini, S. Feng, S. E. Jacobsen, G. E. Robinson, and R. Maleszka. 2012. DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative

phenotypes in honey bees. Proc. Natl. Acad. Sci. U. S. A. 109:4968–73.
Foret, S., R. Kucharski, Y. Pittelkow, G. A. Lockett, and R. Maleszka. 2009. Epigenetic regulation of the honey bee transcriptome: unravelling the nature of methylated genes. BMC Genomics 10:472.
Fraley, C., and A. E. Raftery. 2007. Model-based Methods of Classification : Using the mclust Software in Chemometrics. J. Stat. Softw. 18:1–13.
Francioli, L. C., P. P. Polak, A. Koren, A. Menelaou, S. Chun, I. Renkens, C. M. van Duijn, M. Swertz, C. Wijmenga, G. van Ommen, P. E. Slagboom, D. I. Boomsma, K. Ye, V. Guryev, P. F. Arndt, W. P. Kloosterman, P. I. W. de Bakker, and S. R. Sunyaev. 2015. Genome-wide patterns and properties of de novo mutations in humans. Nat. Genet. 47:822–826. Nature Publishing Group.
Gavery, M. R., and S. B. Roberts. 2013. Predominant intragenic methylation is associated with gene expression characteristics in a bivalve mollusc. PeerJ 1:e215.
Gavery, M., and S. Roberts. 2010. DNA methylation patterns provide insight into epigenetic regulation in the Pacific oyster (Crassostrea gigas). BMC Genomics 11:483.
Gonzalez-Ibeas, D., J. Blanca, C. Roig, M. González-To, B. Picó, V. Truniger, P. Gómez, W. Deleu, A. Caño-Delgado, P. Arús, F. Nuez, J. Garcia-Mas, P. Puigdomènech, and M. A. Aranda. 2007. MELOGEN: an EST database for melon functional genomics. BMC Genomics 8:306.
Hadfield, J. D. 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. J. Stat. Softw. 33:1–22.
Harris, R. A., T. Wang, C. Coarfa, R. P. Nagarajan, C. Hong, S. L. Downey, B. E. Johnson, S. D. Fouse, A. Delaney, Y. Zhao, A. Olshen, T. Ballinger, X. Zhou, K. J. Forsberg, J. Gu, L. Echipare, H. O'Geen, R. Lister, M. Pelizzola, Y. Xi, C. B. Epstein, B. E. Bernstein, R. D. Hawkins, B. Ren, W.-Y. Chung, H. Gu, C. Bock, A. Gnirke, M. Q. Zhang, D. Haussler, J. R. Ecker, W. Li, P. J. Farnham, R. a Waterland, A. Meissner, M. a Marra, M. Hirst, A. Milosavljevic, and J. F. Costello. 2010. Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat. Biotechnol. 28:1097–105.
Hastings, K. E. 1996. Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families. J. Mol. Evol. 42:631–640.
Heim, C., and E. B. Binder. 2012. Current research trends in early life stress and depression: review of human studies on sensitive periods, gene-environment interactions, and epigenetics. Exp. Neurol. 233:102–11. Elsevier Inc.
Herman, J. J., S. E. Sultan, and J. J. Herman. 2016. DNA methylation mediates genetic variation for adaptive transgenerational plasticity.
Hoegh-Guldberg, O., L. Hughes, S. McIntyre, D. B. Lindenmayer, C. Parmesan, H. P. Possingham, and C. D. Thomas. 2008. Assisted colonization and rapid climate change. Science (80-.). 321:345–346.
Hoegh-Guldberg, O., P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira, N. Knowlton, C. M. Eakin, R. Iglesias-Prieto, N. Muthiga, R. H. Bradbury, A. Dubi, and M. E. Hatziolos. 2007. Coral Reefs Under Rapid Climate Change and Ocean Acidification. Science (80-.). 318:1737–1742.
Hoekstra, L. A., M. A. Siddiq, and K. L. Montooth. 2013. Pleiotropic effects of a mitochondrial-nuclear incompatibility depend upon the accelerating effect of temperature in Drosophila. Genetics 195:1129–39.

Hofmann, G. E. 2017. Ecological Epigenetics in Marine Metazoans. Front. Mar. Sci. 4:1–7.
Hughes, T. P., A. H. Baird, D. R. Bellwood, M. Card, S. R. Connolly, C. Folke, R. Grosberg, O. Hoegh-Guldberg, J. B. C. Jackson, J. Kleypas, J. M. Lough, P. Marshall, M. Nystrom, S. R. Palumbi, J. M. Pandolfi, B. Rosen, and J. Roughgarden. 2003. Climate change, human impacts, and the resilience of coral reefs. Science (80-.). 301:929–933.
Huh, I., J. Zeng, T. Park, and S. V Yi. 2013. DNA methylation and transcriptional noise. Epigenetics Chromatin 6:9.
Hunt, B., and J. Brisson. 2010. Functional Conservation of DNA Methylation in the Pea Aphid and the Honeybee. Genome Biol. Evol. 2:719–728.
Ikemura, T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. 151:389–409.
Ingvarsson, P. K. 2001. Restoration of genetic variation lost – the genetic rescue hypothesis. Trends Ecol. Evol. 16:62–63.
Jaenisch, R., and A. Bird. 2003. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat. Genet. Suppl. 33:245–54.
Javierre, B. M., A. F. Fernandez, J. Richter, F. Al-Shahrour, J. I. Martin-Subero, J. Rodriguez-Ubreva, M. Berdasco, M. F. Fraga, T. P. O'Hanlon, L. G. Rider, F. V Jacinto, F. J. Lopez-Longo, J. Dopazo, M. Forn, M. a Peinado, L. Carreño, A. H. Sawalha, J. B. Harley, R. Siebert, M. Esteller, F. W. Miller, and E. Ballestar. 2010. Changes in the pattern of DNA methylation associate with twin discordance in systemic lupus erythematosus. Genome Res. 20:170–9.
Jjingo, D., A. B. Conley, S. V Yi, V. V Lunyak, and I. King. 2012. On the presence and role of human gene-body DNA methylation. Oncotarget 3:462–474.
Jombart, T., S. Devillard, and F. Balloux. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.
Jones, P. A. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat. Rev. Genet. 13:484–92. Nature Publishing Group.
Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura. 2001. Codon usage and tRNA genes in eukaryotes: Correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J. Mol. Evol. 53:290–298.
Karlin, S., and J. Mrázek. 1996. What drives codon choices in human genes? J. Mol. Biol. 262:459–72.
Katoh, K., and D. M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.
Keller, T. E., P. Han, and S. V Yi. 2015. Evolutionary Transition of Promoter and Gene Body DNA Methylation across Invertebrate-Vertebrate Boundary. Mol. Biol. Evol. 33:1–30.
Kelly, S. A., T. M. Panhuis, and A. M. Stoehr. 2012. Phenotypic plasticity: molecular mechanisms and adaptive significance. Compr. Physiol. 2:1417–39.
Kenkel, C. D., E. Meyer, and M. V Matz. 2013. Gene expression under chronic heat stress in populations of the mustard hill coral (Porites astreoides) from different thermal environments. Mol. Ecol. 22:4322–34.
Kenkel, C., and M. V Matz. 2016. Enhanced gene expression plasticity as a mechanism of adaptation to a variable environment in a reef-building coral. Nat. Ecol. Evol. 1:14. Nature Publishing Group.
Kenyon, J. C. 1997. Models of reticulate evolution in the coral genus Acropora based on

chromosome numbers: Parallels with plants. Evolution (N. Y). 51:756–767.
Kitahara, M. V, S. D. Cairns, J. Stolarski, D. Blair, and D. J. Miller. 2010. A comprehensive phylogenetic analysis of the Scleractinia (Cnidaria, Anthozoa) based on mitochondrial CO1 sequence data. PLoS One 5:e11490.
Kitchen, S. A., C. M. Crowder, A. Z. Poole, V. M. Weis, and E. Meyer. 2015. De novo assembly and characterization of four anthozoan (phylum Cnidaria) transcriptomes. G3 Genes Genomes Genet. 5:2441–2452.
Krueger, F., and S. R. Andrews. 2011. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27:1571–1572.
Kucharski, R., J. Maleszka, S. Foret, and R. Maleszka. 2008. Nutritional control of reproductive status in honeybees via DNA methylation. Science (80-.). 319:1827–30.
Langfelder, P., and S. Horvath. 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559.
Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–9.
Law, J. a, and S. E. Jacobsen. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat. Rev. Genet. 11:204–220. Nature Publishing Group.
Lehnert, E. M., M. E. Mouchka, M. S. Burriesci, N. D. Gallo, J. a Schwarz, and J. R. Pringle. 2014. Extensive differences in gene expression between symbiotic and aposymbiotic cnidarians. G3 4:277–95.
Levy, O., P. Kaniewska, S. Alon, E. Eisenberg, S. Karako-Lampert, L. K. Bay, R. Reef, M. Rodriguez-Lanetty, D. J. Miller, and O. Hoegh-Guldberg. 2011. Complex diel cycles of gene expression in coral-algal symbiosis. Science (80-.). 331:175.
Libbrecht, R., P. R. Oxley, L. Keller, and D. J. C. Kronauer. 2016. Robust DNA methylation in the clonal raider ant brain. Curr. Biol. 26:391–395. Elsevier Ltd.
Logan, C. A., J. P. Dunne, C. M. Eakin, and S. D. Donner. 2014. Incorporating adaptive responses into future projections of coral bleaching. Glob. Chang. Biol. 20:125–139.
Lohman, B. K., J. N. Weber, and D. I. Bolnick. 2016. Evaluation of TagSeq, a reliable low-cost alternative for RNAseq. Mol. Ecol. Resour. 16:1315–1321.
Love, M. I., W. Huber, and S. Anders. 2014. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol. 15:1–21.
Lubinski, T., and B. Granger. 2013. PocilloporaBase.
Lyko, F., S. Foret, R. Kucharski, and S. Wolf. 2011. The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol. 8.
Maor-Landaw, K., S. Karako-Lampert, H. W. Ben-Asher, S. Goffredo, G. Falini, Z. Dubinsky, and O. Levy. 2014. Gene Expression Profiles during short-term heat stress in the Red Sea Coral Stylophora pistillata. Glob. Chang. Biol., doi: 10.1111/gcb.12592.
Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 17:10–12.
Maunakea, A. K., R. P. Nagarajan, M. Bilenky, T. J. Ballinger, C. D. Souza, S. D. Fouse, B. E. Johnson, C. Hong, C. Nielsen, Y. Zhao, G. Turecki, A. Delaney, R. Varhol, N. Thiessen, K. Shchors, V. M. Heine, D. H. Rowitch, X. Xing, C. Fiore, M. Schillebeeckx, S. J. M. Jones, D. Haussler, M. A. Marra, M. Hirst, T. Wang, and J. F. Costello. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466:253–257. Nature Publishing Group.
McGaughey, D. M., H. O. Abaan, R. M. Miller, P. a Kropp, and L. C. Brody. 2014. Genomics of

CpG methylation in developing and developed zebrafish. G3 Genes Genomes Genet. 4:861–9.
McKenna, A., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20:1297–303.
Meyer, E., G. V Aglyamova, and M. V Matz. 2011. Profiling gene expression responses of coral larvae (Acropora millepora) to elevated temperature and settlement inducers using a novel RNA-Seq procedure. Mol. Ecol. 20:3599–616.
Meyer, E., S. Davies, S. Wang, B. Willis, D. Abrego, T. Juenger, and M. Matz. 2009. Genetic variation in responses to a settlement cue and elevated temperature in the reef-building coral Acropora millepora. Mar. Ecol. Prog. Ser. 392:81–92.
Miyake, K., C. Yang, Y. Minakuchi, K. Ohori, M. Soutome, T. Hirasawa, Y. Kazuki, N. Adachi, S. Suzuki, M. Itoh, Y.-I. Goto, T. Andoh, H. Kurosawa, M. Oshimura, M. Sasaki, A. Toyoda, and T. Kubota. 2013. Comparison of Genomic and Epigenomic Expression in Monozygotic Twins Discordant for Rett Syndrome. PLoS One 8:e66729.
Moya, A., L. Huisman, E. E. Ball, D. C. Hayward, L. C. Grasso, C. M. Chua, H. N. Woo, J. P. Gattuso, S. For??t, and D. J. Miller. 2012. Whole Transcriptome Analysis of the Coral Acropora millepora Reveals Complex Responses to CO 2-driven Acidification during the Initiation of Calcification. Mol. Ecol. 21:2440–2454.
Murgatroyd, C., A. V Patchev, Y. Wu, V. Micale, Y. Bockmühl, D. Fischer, F. Holsboer, C. T. Wotjak, O. F. X. Almeida, and D. Spengler. 2009. Dynamic DNA methylation programs persistent adverse effects of early-life stress. Nat. Neurosci. 12:1559–66.
Newman, J. R. S., S. Ghaemmaghami, J. Ihmels, D. K. Breslow, M. Noble, J. L. DeRisi, and J. S. Weissman. 2006. Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise. Nature 441:840–6.
Nicotra, A. B., O. K. Atkin, S. P. Bonser, A. M. Davidson, E. J. Finnegan, U. Mathesius, P. Poot, M. D. Purugganan, C. L. Richards, F. Valladares, and M. Van Kleunen. 2010. Plant phenotypic plasticity in a changing climate. Trends Plant Sci. 15:684–692. Elsevier Ltd.
Nordberg, H., M. Cantor, S. Dusheyko, S. Hua, A. Poliakov, I. Shabalov, T. Smirnova, I. V. Grigoriev, and I. Dubchak. 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Res. 42:26–31.
Ooijen, J. W. Van. 2006. JoinMap 4 ®.
Otto, S., and T. Day. 2007. A Biologist's Guide to Mathematical Modeling in Ecology and Evolution. Princeton University Press, Princeton, NJ.
Pál, C., B. Papp, and L. D. Hurst. 2001. Highly Expressed Genes in Yeast Evolve Slowly. Genetics 158:927–931.
Palumbi, S. R., D. J. Barshis, N. Traylor-Knowles, and R. A. Bay. 2014. Mechanisms of Reef Coral Resistance to Future Climate Change. Science (80-.). 344:895–898.
Park, J., Z. Peng, J. Zeng, N. Elango, T. Park, D. Wheeler, J. H. Werren, and S. V Yi. 2011. Comparative Analyses of DNA Methylation and Sequence Evolution Using Nasonia Genomes. Mol. Biol. Evol. 28:3345–3354.
Peden, J. F. 1999. Analysis of codon usage. Thesis Submitt. to Univ. Nottingham 50–90.
Plotkin, J. B., and G. Kudla. 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12:32–42. Nature Publishing Group.
Polato, N. R., J. C. Vera, and I. B. Baums. 2011. Gene discovery in the threatened elkhorn coral:

454 sequencing of the acropora palmata transcriptome. PLoS One 6:e28634.
Portela, A., and M. Esteller. 2010. Epigenetic modifications and human disease. Nat. Biotechnol. 28:1057–68.
Price, T. D., A. Qvarnström, and D. E. Irwin. 2003. The role of phenotypic plasticity in driving genetic evolution. Proc. R. Soc. London 270:1433–1440.
Putnam, H. M., J. Davidson, and R. D. Gates. 2016. Ocean acidification influences DNA methylation and phenotypic plasticity in environmentally susceptible corals. Integr. Comp. Biol. 56:E177–E177.
Putnam, N. H., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, A. Terry, H. Shapiro, E. Lindquist, V. V Kapitonov, J. Jurka, G. Genikhovich, I. V Grigoriev, S. M. Lucas, R. E. Steele, J. R. Finnerty, U. Technau, M. Q. Martindale, and D. S. Rokhsar. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94.
Qin, Z., Z. Cai, G. Xia, and M. Wang. 2013. Synonymous codon usage bias is correlative to intron number and shows disequilibrium among exons in plants. BMC Genomics 14:56. BMC Genomics.
Quinlan, A. R., and I. M. Hall. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
R Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Australia.
Reusch, T. B. H. 2013. Climate change in the oceans : evolutionary versus phenotypically plastic responses of marine animals and plants. Evol. Appl., doi: 10.1111/eva.12109.
Richards, E. J. 2008. Population epigenetics. Curr. Opin. Genet. Dev. 18:221–226.
Riviere, G., G.-C. Wu, A. Fellous, D. Goux, P. Sourdaine, and P. Favrel. 2013. DNA methylation is crucial for the early development in the Oyster C. gigas. Mar. Biotechnol. (NY). 15:739–53.
Roberts, S. B., and M. R. Gavery. 2012. Is there a relationship between DNA methylation and phenotypic plasticity in invertebrates? Front. Physiol. 2:1–5.
Rocker, M., B. Willis, and L. Bay. 2012. Thermal stress-related gene expression in corals with different Symbiodinium types. Proc. 12th Int. Coral Reef Symp. 9:1–5.
Rohner, N., D. F. Jarosz, J. E. Kowalko, M. Yoshizawa, W. R. Jeffery, R. L. Borowsky, S. Lindquist, and C. J. Tabin. 2013. Cryptic variation in morphological evolution: HSP90 as a capacitor for loss of eyes in cavefish. Science 342:1372–5.
Romano, S. L., and S. R. Palumbi. 1996. Evolution of Scleractinian Corals Inferred from Molecular Systematics. Science (80-.). 271:640–642.
Rumble, S. M., P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno. 2009. SHRiMP: Accurate mapping of short color-space reads. PLoS Comput. Biol. 5:1–11.
Sarda, S., J. Zeng, B. G. Hunt, and S. V. Yi. 2012. The evolution of invertebrate gene body methylation. Mol. Biol. Evol. 29:1907–1916.
Schwarz, J. a, P. B. Brokstein, C. Voolstra, A. Y. Terry, C. F. Manohar, D. J. Miller, A. M. Szmant, M. A. Coffroth, and M. Medina. 2008. Coral life history and symbiosis: functional genomic resources for two reef building Caribbean corals, Acropora palmata and Montastraea faveolata. BMC Genomics 9:97.
Seneca, F. O., and S. R. Palumbi. 2015. The role of transcriptome resilience in resistance of corals to bleaching. Mol. Ecol. 24:1467–1484.
Sharp, P. M., and W. H. Li. 1987a. The codon adaptation index-a measure of directional

synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.
Sharp, P. M., and W. H. Li. 1987b. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4:222–230.
Sharp, P. M., T. M. F. Tuohy, and K. R. Mosurski. 1986. Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. 14:5125–5143.
Shen, J. C., W. M. Rideout, and P. a Jones. 1994. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. 22:972–976.
Shinzato, C., E. Shoguchi, T. Kawashima, M. Hamada, K. Hisata, M. Tanaka, M. Fujie, M. Fujiwara, R. Koyanagi, T. Ikuta, A. Fujiyama, D. J. Miller, and N. Satoh. 2011. Using the Acropora digitifera genome to understand coral responses to environmental change. Nature 476:320–3. Nature Publishing Group.
Stam, P. 1993. Construction of integrated genetic linkage maps by means of a new computer package : JOINMAP. Plant J. 3:739–744.
Stamatakis, A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Sterky, F., R. R. Bhalerao, P. Unneberg, B. Segerman, P. Nilsson, A. M. Brunner, L. Charbonnel-Campaa, J. J. Lindvall, K. Tandre, S. H. Strauss, B. Sundberg, P. Gustafsson, M. Uhlén, R. P. Bhalerao, O. Nilsson, G. Sandberg, J. Karlsson, J. Lundeberg, and S. Jansson. 2004. A Populus EST resource for plant functional genomics. Proc. Natl. Acad. Sci. U. S. A. 101:13951–13956.
Subramanian, S., and S. Kumar. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373–381.
Sun, J., Q. Chen, J. C. Y. Lun, J. Xu, and J. W. Qiu. 2013. PcarnBase: Development of a Transcriptomic Database for the Brain Coral Platygyra carnosus. Mar. Biotechnol. 15:244–251.
Sunagawa, S., E. C. Wilson, M. Thaler, M. L. Smith, C. Caruso, J. R. Pringle, V. M. Weis, M. Medina, and J. a Schwarz. 2009. Generation and analysis of transcriptomic resources for a model system on the rise: the sea anemone Aiptasia pallida and its dinoflagellate endosymbiont. BMC Genomics 10:258.
Suyama, M., D. Torrents, and P. Bork. 2006. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:609–612.
Suzuki, M. M., and A. Bird. 2008. DNA methylation landscapes: provocative insights from epigenomics. Nat. Rev. Genet. 9:465–76.
Suzuki, M. M., A. R. W. Kerr, D. De Sousa, and A. Bird. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. Genome Res. 17:625–631.
Suzuki, M. M., A. Yoshinari, M. Obara, S. Takuno, S. Shigenobu, Y. Sasakura, A. R. Kerr, S. Webb, A. Bird, and A. Nakayama. 2013. Identical sets of methylated and nonmethylated genes in Ciona intestinalis sperm and muscle cells. Epigenetics Chromatin 6:38.
Sved, J., and a Bird. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl. Acad. Sci. U. S. A. 87:4692–6.
Takuno, S., and B. S. Gaut. 2012. Body-methylated genes in arabidopsis thaliana are functionally important and evolve slowly. Mol. Biol. Evol. 29:219–227.
Takuno, S., J.-H. Ran, and B. S. Gaut. 2016. Evolutionary patterns of genic DNA methylation vary across land plants. Nat. Plants 2:15222. Nature Publishing Group.
Tran, R. K., J. G. Henikoff, D. Zilberman, R. F. Ditt, S. E. Jacobsen, and S. Henikoff. 2005.

DNA methylation profiling identifies CG methylation clusters in Arabidopsis genes. Curr. Biol. 15:154–159.
Tucker, K. L. 2001. Methylated cytosine and the brain: a new base for neuroscience. Neuron 30:649–52.
Van Oppen, M. J. H., L. M. Peplow, S. Kininmonth, and R. Berkelmans. 2011. Historical and contemporary factors shape the population genetic structure of the broadcast spawning coral, Acropora millepora, on the Great Barrier Reef. Mol. Ecol. 20:4899–914.
Verhoeven, K. J. F., B. M. VonHoldt, and V. L. Sork. 2016. Epigenetics in ecology and evolution: What we know and what we need to know. Mol. Ecol. 25:1631–1638.
Wang, H., G. Beyene, J. Zhai, S. Feng, N. Fahlgren, N. J. Taylor, R. Bart, J. C. Carrington, S. E. Jacobsen, and I. Ausin. 2015. CG gene body DNA methylation changes and evolution of duplicated genes in cassava. Proc. Natl. Acad. Sci. U. S. A. 112:13729–13734.
Wang, S., E. Meyer, J. K. McKay, and M. V Matz. 2012. 2b-RAD: a simple and flexible method for genome-wide genotyping. Nat. Methods 9:808–810.
Wang, W.-S., Y.-J. Pan, X.-Q. Zhao, D. Dwivedi, L.-H. Zhu, J. Ali, B.-Y. Fu, and Z.-K. Li. 2011. Drought-induced site-specific DNA methylation and its association with drought tolerance in rice (Oryza sativa L.). J. Exp. Bot. 62:1951–60.
Wang, X., Q. Li, J. Lian, L. Li, L. Jin, H. Cai, F. Xu, H. Qi, and L. Zhang. 2014. Genome-wide and single-base resolution DNA methylomes of the Pacific oyster Crassostrea gigas provide insight into the evolution of invertebrate CpG methylation. BMC Genomics 15.
Wang, X., J. H. Werren, and A. G. Clark. 2016. Allele-Specific Transcriptome and Methylome Analysis Reveals Stable Inheritance and Cis-Regulation of DNA Methylation in Nasonia. PLOS Biol. 14:e1002500.
Wang, X., D. Wheeler, A. Avery, A. Rago, J.-H. Choi, J. K. Colbourne, A. G. Clark, and J. H. Werren. 2013. Function and evolution of DNA methylation in Nasonia vitripennis. PLoS Genet. 9:e1003872.
Whitlock, M. C. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. J. Evol. Biol. 18:1368–73.
Willette, D. a., F. W. Allendorf, P. H. Barber, D. J. Barshis, K. E. Carpenter, E. D. Crandall, W. a. Cresko, I. Fernandez-Silva, M. V. Matz, E. Meyer, M. D. Santos, L. W. Seeb, and J. E. Seeb. 2014. So, you want to use next-generation sequencing in marine systems? Insight from the Pan-Pacific Advanced Studies Institute. Bull. Mar. Sci. 90:79–122.
Wright, F. 1990. The "effective number of codons" used in a gene. Gene 87:23–29.
Xiang, H., J. Zhu, Q. Chen, F. Dai, X. Li, M. Li, H. Zhang, G. Zhang, D. Li, Y. Dong, L. Zhao, Y. Lin, D. Cheng, J. Yu, J. Sun, X. Zhou, K. Ma, Y. He, Y. Zhao, S. Guo, M. Ye, G. Guo, Y. Li, R. Li, X. Zhang, L. Ma, K. Kristiansen, Q. Guo, J. Jiang, S. Beck, Q. Xia, W. Wang, and J. Wang. 2010. Single base-resolution methylome of the silkworm reveals a sparse epigenomic map. Nat. Biotechnol. 28:516–20. Nature Publishing Group.
Yang, L., and B. S. Gaut. 2011. Factors that contribute to variation in evolutionary rate among arabidopsis genes. Mol. Biol. Evol. 28:2359–2369.
Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.
Yeh, P. J., and T. D. Price. 2004. Adaptive phenotypic plasticity and the successful colonization of a novel environment. Am. Nat. 164:531–42.
Zemach, A., I. E. McDaniel, P. Silva, and D. Zilberman. 2010. Genome-wide evolutionary analysis of eukaryotic DNA methylation. Science (80-.). 328:916–9.

Zemach, A., and D. Zilberman. 2010. Evolution of Eukaryotic DNA Methylation and the Pursuit of Safer Sex. Curr. Biol. 20:R780–R785. Elsevier Ltd.

Zhang, L., and W. H. Li. 2004. Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. Mol. Biol. Evol. 21:236–239.

Zhang, X., J. Yazaki, A. Sundaresan, S. Cokus, S. W. L. Chan, H. Chen, I. R. Henderson, P. Shinn, M. Pellegrini, S. E. Jacobsen, and J. Ecker. 2006. Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. Cell 126:1189–1201.

Zilberman, D. 2017. An evolutionary case for functional gene body methylation in plants and animals. Genome Biol. 18:87. Genome Biology.

Zilberman, D., M. Gehring, R. K. Tran, T. Ballinger, and S. Henikoff. 2007. Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat. Genet. 39:61–69.